The effect of Open Data on the software development process: an exploratory approach

Master's Thesis - Human Centered Informatics

Augustin Le Fèvre – June 2014

101 347 characters
42 pages

Supervised by Anne Marie Kanstrup

Abstract

Open Data (data publicly accessed that anyone is free to use, reuse and redistribute) has been studied as a mean to foster economic growth, to promote more transparent governance, or to empower the citizens of a country. These studies do not analyse the Open Data's impact on individuals and companies developing software, or Application Providers. Thus, this thesis seeks to fill this gap by studying the effects of Open Data on the Application Providers' software development methods and on the ecosystem they are part of it. In order to answer these questions, this thesis applies an exploratory approach model as its main methodology. This implies that the goal of this research is not to confirm a priori hypotheses but to make an initial contribution to the study about the impact of Open Data on Application Providers and to propose avenues for possible future researches.

The first step of this study was to explore the literature to create an initial knowledge about the topic, in order to determine what an analysis of this topic should enquire. The research also included a data creation process, which consists of four successive semi-structured interviews in an iterative process: the outcome of an interview contributed to the redefinition of the interview guide of the next one. The collected data was analysed, working through codes, categories and finally themes while building a higher level of data understanding of the data. Regarding the analysis, it was possible to create a model of the development process, identifying new tasks and explaining the way some common tasks need to be addressed when working with Open Data. Firstly, after identification of a need and finding a way to solve the problem, the Application Providers have to retrieve data. Then they have to assure the reliability of this data (a certification task), to clean/correct the data, and finally to create a wrapper around the data to render the uses easier, before being able to develop an application. Afterwards this thesis provided with an explanation why and how beneficial it is for an Application Provider to have employees with certain skills. These skills are the ones of a journalist (who knows how to look for and understand data), data specialist (who can process and treat data efficiently), XML expert (who knows how to create models of complex data relations) and designer. Finally, the thesis explained how the fact that Application Providers had opened their data both internally and externally improved their capacity to innovate. Lastly, this thesis described how some of the tasks mentioned previously could be outsourced by the Application Provider, bringing a new actor in the ecosystem, who proposes the retrieval and treatment of the data as a service.

Acknowledgments

I would like to thank my supervisor Anne Marie Kanstrup who supported me and encouraged me during the creation of my Master's thesis. I would also like to thank Ellen Christiansen and Pär-Ola Zander for their observations and suggestions during the seminars they organised for the Master's students. I am grateful for my classmates for their support, especially Simon Kristoffer Johansen for this long discussion early in the semester. I would also like to thank Arnaud Angelo, Paul Aubineau, Marc Brice and Christophe Desclaux (and Reador.NET). Without them, I could not have written this result chapter. Lastly, a special thanks to my family, to my friends and to those who helped me in some way or another.

Key technical words

Since the reader of this thesis might not be familiar with the software development vocabulary, I consider that I have to introduce three expressions that I will use later in this thesis.

API: This acronym stands for Application Programmable Interface and consists of all kind of interfaces letting a developer to access to some software. Whereas an API can be used to access the functionalities of some specific library, in this project, I will mostly use it to designate web API, i.e. the way companies open an access to their data.

Hackathon: A hackathon (coined from hack and marathon) is an event where people involved in software development (developers, graphic designers and project managers) create a working application in a limited amount of time (usually 24 or 48 hours). They are usually grouped in teams competing for a recompense. Most of these hackathons have a central theme (usually defined by the organiser/main sponsor), such as a specific technology, a social consideration ("applications designed to help diabetic people"), or some specific datasets/APIs.

REST/RESTful: REST is a technical standard developed to describe how a web API should be designed. Developers identified several level of REST (an API can be more or less REST), an API reaching the maximum level (third one) of REST is considered RESTful.

Table of Contents

Introduction

The technological development is one of the main factors fuelling today's world. While becoming an essential part of many different fields of people's lives it contributes to improvement of living standards, economic growth, and an access to data. Today almost everyone who possesses one or another type of electronic device is able to obtain information he needs in a matter of seconds, therefore, a free and open access to different types of data became a need for many people. A data or content is open if "anyone is free to use, reuse and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike" (Open Definition, 2005). Open Data is the consequence of a citizen-driven movement. The Open Data (OD) movement is rather new and promotes the openness of data, through non-governmental organisations such as the Open Knowledge Foundation (OKFN). There are several objectives behind OD. Indeed, those organisations promoting it believe that OD will bring better governance, culture, research and economy (Open Knowledge Foundation, 2014).

History

Before presenting OD, a brief history of how it started and of its main actors is beneficial for the following of the project. The majority of the discussions and actions promoting OD started around 2004. The science ministers of the OECD (Organisation for Economic Co-operation and Development) signed a declaration encouraging the openness of data publicly funded (OECD, 2004). The OKFN was also founded in 2004. During the first years, most of the OD was produced by private people through organisations, such as Wikipedia, or Open Street Map (OSM), or came from scientific publications (as promoted by the OECD). However, the big turning point was the signature by Barrack Obama on his first day in office, of the *Memorandum on Transparency and Open Government*, which promoted the openness of governmental produced data (Obama, 2009). In May of the same year, the U.S. government launched the data.gov website, which now hosts more than 85,000 datasets (data.gov, 2014).

The data

Now that I provided the reader with a context, it is fundamental to define the nature of such data. As explained earlier, there are two main types of data. The first one is crowdsourced data (such as OSM), which usually aims at building repositories of specific knowledge (maps, encyclopaedia, etc.). The second one is the Public Sector Information (PSI). An EU report, the

MESPIR report considers that PSI represents the single largest source of information in Europe and states that its overall market size (not the value of the data, but what could be done with it) ranges from 10 to 48 billion euros (MESPIR, 2006). PSI can be divided in six main categories (Alexopoulos, Spiliotopoulou, & Charalabidis, 2013):

- *Business information:* this includes information related to the public administrations referencing businesses, such as Chamber of Commerce, or patent and trademark information;
- *Geographic information:* this includes data necessary to draw maps, such as "address information, aerial photos, buildings, cadastral information, geodetic networks, geology, hydrographical data and topographic information";
- *Legal information:* this includes every kind of legislation in the country, both national and dependent on treaties or foreign courts (such as the *Court of Justice of the European Union* in the EU);
- *Meteorological information:* this includes information related to the weather, from data and models to weather forecasts;
- *Social data:* this includes statistics realised by the administrations such as "economic, public spending, employment, health, population, public administration, social";
- *Transport information:* this includes any information helping to commute in the country, from information on the public transport systems (schedules, maps), information on the roads, etc.

Whereas these six categories obviously do not cover all the possible types of PSI, they do represent a large amount of data, cover many different situations and as pointed in the MESPIR report, they have a significant market size.

Open Data sources

After explaining the nature of data, and making the distinction between crowdsourced data and PSI, I will give more details of the way PSI is accessible. PSI, when opened by a government, is supposed to be hosted online. The ENGAGE project, a project funded by the European Union, which aims at the "deployment and use of an advanced service infrastructure, incorporating distributed and diverse public sector information resources as well as data curation, semantic annotation and visualization tools, capable of supporting scientific collaboration and governance-related research from multidisciplinary scientific communities, while also

empowering the deployment of open governmental data towards citizens" (ENGAGE, 2014), has studied those sources.

After studying many different sources, the authors of this project identified three different categories of sources (ENGAGE, 2011):

- *PSI Catalogues / Open Government Data Catalogue Aggregators:* those are portals that "publish and maintain lists of PSI catalogues". Whereas they do not provide direct links to data, they aim to make the search for data repositories easier.
- *Public Sector Sources and Open Data initiatives:* those are the portals that initially store a dataset. This category includes most of the governmental portals (data.gov, data.gov.uk, data.gouv.fr etc.). Usually, data is handled on these portals from its publication, to its update, etc. Usually, these portals (or the organisations running them) are the licensers of the data (ideally, these licenses are open ones, such as the Open Database License, edited by the OKFN and which follows the Open Definition principles). It is also common to find example of uses of the datasets, link to applications using them, etc.
- *Web-based platforms for sharing Data Sources URLs:* this category includes platforms that also link to sources, but in a more structured way, with a better focus on the datasets. They are not limited to one source of data. This includes important platforms such as the CKAN data hub portal, managed by the OKFN.

Whereas these websites offer the same kind of information, they do not necessarily offer it in the same way. Some follow a "Web 1.0 paradigm" by being more traditional in their approach and other follow a "Web 2.0 paradigm", by having a more advanced infrastructure (Charalampos, Loukis, Charalabidis, & Zuiderwijk, 2013).

Traditional platforms are simple repositories of datasets. They do not provide a good service to users. Datasets are often in non-machine-processable format (PDF files for instance). Most of the time, they do not provide any details about the datasets, or information to link datasets between them. They also allow the users to modify or "clean" the data. Indeed, those datasets, while coming from trusted sources, are not exempt of mistakes. Their format might not be optimal or there might be duplications between datasets.

On the other hand, there are more advanced platforms, which really embody the notion of openness. They benefit from the advanced possibilities offered by the Information and Communication Technologies (ICT) to increase the efficiency of such platforms. In this situation, instead of being simply a passive consumer of content, the user can also contribute to

the content. This happens through different actions, from actively working on the data (cleaning datasets, noting those that are duplicates, creating different versions for special needs), to improve the information related to the datasets, with comments, rates, or meta-data. With this process, we can start to speak of "pro-sumer" (Charalampos, Loukis, Charalabidis, & Zuiderwijk, 2013) instead of simple consumers. People searching for content benefit from these platforms in many ways. They are able to find more easily what they are looking for, to know about the quality of datasets founds, to find datasets of a better quality and to increase their ability to link the different datasets related to the topic they are investigating.

Linked data

In this section, I will speak of Linked Data (LD), a publishing format for data. This is relevant to this thesis as. While OD is not necessarily linked, it can be and it benefits from being so.

What is Linked Data

LD is a term coined by Tim Berners-Lee, co-founder and president of the World Wide Web Consortium (W3C). He is also a co-founder and the president of the Open Data Institute (ODI), one of the major organisations working on OD. In 2006, he presented a paper arguing that putting data on the web is only a first step. What is necessary to provide is a context to this data by linking it to other people's data (Bernes-Lee, 2006).

LD as a data publishing method, as the name already indicates, is one of the methods used to link data to other people's data. This has some rather technical rules, however, the concept has been summarised in three principles that Tim Berners-Lee presented at a TED conference (Berners-Lee, 2009):

- "All kinds of conceptual things, they have names now that start with HTTP."
- "I get important information back. I will get back some data in a standard format which is kind of useful data that somebody might like to know about that thing, about that event."
- "I get back that information it's not just got somebody's height and weight and when they were born, it's got relationships. And when it has relationships, whenever it expresses a relationship then the other thing that it's related to is given one of those names that starts with HTTP."

Linked and Open Data

I mentioned before that linking datasets together improves their quality, but I did not explained why. I was using the term "link" without referring to the "Linked Data" standard. However, the standard that was introduced in the last paragraphs is actually considered as a way to publish OD. The ODI explains that publishing OD as LD has two main benefits, which are (ODI, 2014):

- being able to link up and merge together data from different sources that refer to the same identified thing;
- being able to publish data in a distributed way, with intermediaries collecting and aggregating data.

It is worth mentioning that we can find a scheme in Tim Berners-Lee's article, which he added in 2010, to know if the LD is good or not, and this scheme is actually very close to the one used by the OKFN in their index of countries:

- Available on the web (whatever format) but with an open licence, to be Open Data;
- Available as machine-readable structured data (e.g. excel instead of image scan of a table);
- Available in a non-proprietary format (e.g. CSV instead of excel);
- All the above plus, use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff;
- All the above, plus: Link your data to other people's data to provide context.

Current state of the Open Data

The OKFN, through its *Open Data Index* publishes a crowdsourced index of the data opened by the different governments. It lists ten types of essential datasets (Open Knowledge Foundation, 2014):

- Transport Timetables;
- Government Budget;
- Government Spending;
- Election Results;
- Company Register;
- National Map;
- National Statistics;

- Legislation;
- Postcodes / Zip codes;
- Emissions of pollutants;

For each of the seventy countries indexed, people look if such datasets:

- Exists;
- Are digital;
- Are publicly available;
- Are free of charge;
- Are online;
- Are machine readable;
- Are available in bulk;
- Are openly licensed;
- Are up-to-date;

On this portal, a total of 700 datasets are identified, and 84 of them are open (Open Knowledge Foundation, 2014). In this sense, the most open country is the UK, which reaches a score of 940, having all these datasets open, with some minor problems (for instance, the election results are not openly licensed, or the emissions of pollutants are not available in bulk).

We can see that while there is still a lot to be done, governments are seeking for improvements on the subject. On the other hand, there are citizen driven initiatives where people (or organisations) contribute on their own on to the creation of public repositories of knowledge, crowdsourced data. Usually, these databases are managed by non-profit organisations. The most famous of this kind of websites is Wikipedia (an encyclopaedia), which is the only one non-profit website among thirty of the most viewed websites in the world. Another example would be OpenStreetMap, which publishes maps of the whole world. It can be used as a service through their website (like any other online map, Google Maps, Nokia Here, etc.), but the whole content can also be downloaded and then used independently from the website.

Consequences of the existence of Open Data

In this part, I will specify the questions that I am going to answer through this thesis. Firstly, I will present the research that has already been made and then I will determine what my perspective on the subject of OD is. Finally, I will reflect on some possible interrogations related to the topic.

State of the current researches

OD, while being relatively new, has already been subject of several studies. Some studies were made about the relation between OD and democracy. Tim Davies analysed the uses of governmental data from data.gov.uk and the "possible implications that has for different models of democratic change and public sector reform" (Davis, 2010). Michel Gurstein studied if OD was actually used by everyone or just increased the empowerment of the already empowered citizens (Gurstein, 2011).

Other researchers were more interested in its economic consequences. I already pointed at the MESPIR report (MESPIR, 2006) which estimated the value of such data. *The Economist* published an article comparing the freeing of data to the fact that the US made the GPS available to everyone in the world for free. Considering that three millions of US jobs would not exist without the GPS, OD could also foster the creation of jobs (The Economist, 2013).

However, there are way less studies about the relation between OD and the business or software developers. Scholars agreed on its economic impact, but this did not explain how the software industry would have to change in order to work with this new data, i.e. if there is a need to change the software development process.

Perspective

It is possible at different levels to identify four distinct roles when interacting with the data, (Latif, Saeed, Hoefler, Stocker, & Wagner, 2009):

- *A raw data provider* provides datasets not in a RDF format, i.e. without providing meta-data or other data relevant to the dataset;
- *A linked data provider* provides data readable by a machine, in a Linked Data format;
- *A linked data application provider* creates an application, which makes the data readable by a human;
- *An end user* consumes the data through the application.

Many studies focus on how data providers should build their platforms and how they should publish their data, usually identifying the need of meta-data (Zuiderwijk, Jeffery, & Janssen, 2012), or studying the ways to improve it (Eberius, Braunschweig, Thiele, & Lehner, 2012). These studies are especially relevant to providers of raw data or linked data, or to organisations contributing to the OD movement. However, they do not inform us about the situation of the

linked data application providers (which I will call Application Providers or AP). It is the category of people and organisations that I would like to focus on.

Software development methods

We just discussed about the potential influence of OD on the relationships of APs and the community. However, OD could have an impact on software development methods. During the nineties, people started to figure out that the methods they used were not efficient enough for an industry working with constantly changing requirements. Then developers considered working with shorter iterations. Agile methodology was developed (Beck, et al., 2001). I do not claim that OD changes software methodologies as much as agile methods did, however, we could see a real change in these methodologies.

OD is free, contrary to most of closed data, but gathering the right data for a project requires some efforts. Measuring them would be interesting. Moreover, once data is collected, do the following steps of the development process need to be changed? Or are they similar to any other kind of application?

Being part of a community

We notice that many software companies play a part in the Open Source community, contributing to its projects. There is also an OD community and this part will explain why we could imagine a similar relationship between APs and this community similar to the one between these companies and the Open Source community.

It discussed earlier the fact that the new platforms tend to transform data consumers to "pro-sumers" (Charalampos, Loukis, Charalabidis, & Zuiderwijk, 2013). The idea is that the consumers of the data can increase the quality of this data, by contributing to the public repositories. However, this paper, while studying this idea of "pro-sumers", does not focus on what the motivations of the users are, why would they contribute and who actually contributes to it. Since organisations benefit from linked datasets, do they (or should they?) contribute to those, like any other member of the OD community? One of the reasons why I am raising this question is related to the links between software companies and Open Source.

It is quite common for private companies to contribute to the Free and Open Source Software (FOSS) movement. Google, Microsoft, Facebook, Twitter and many others publish this kind of software:

- Google's GitHub account (a platform used by developers to publish and contribute to Open Source projects) has 64 different projects and they even develop an Open Source language (Go);
- Twitter has published and actively maintains a famous front-end framework, called Bootstrap;
- Microsoft has developed a platform called CodePlex to help people to publish Open Source software, moreover, this company actively contribute to it;
- Facebook has around fifty projects on its GitHub account and gave some of their biggest Open Source projects to the Apache Foundation, such as Cassandra (the tenth biggest Database Management System (DB-Engines, 2014)) or Thrift (software used to manage different kinds of servers).

Contributing to the OD movement when using this type of data, could be a possibility, since we can find similarities (which I will explain later) in both movements. Does this mean that the companies building applications on top of OD would have an interest to contribute to the quality of those public datasets and would be "pro-sumers" of this data?

Research question

Through analysis and discussion, this project aims to study the effect of Open Data on Application Providers. Do the Application Providers' software development methods need to be changed to work with such data? How does the existence of Open Data shape the ecosystem around the Application Providers?

Methods

In this chapter, I will present the methods applied to answer my research questions. In general, the methods chapter of a research is one of the "most important part of a scientific paper because it provides the essential information that allows the reader to judge the validity of the results and conclusions of the study reported" (Azevedo, et al., 2011). I will divide this chapter in two parts. The first one presents my methodological approach, which helps to conduct my research as a whole and to decide which methods I should rely on. The second part is a presentation of my data collection process.

Exploratory research

In order to explain the choice of my methods and the way I conduct my research, relating my thesis to a general methodological approach is an important step to do. Indeed, whereas it is not necessary to rigorously follow a procedure, it is helpful to make the appropriate decisions about the methods I choose; moreover, it would help to highlight the weaknesses and strengths of my approach. In this section, I will explain what exploratory research is and why this thesis comes follows an approach close from an exploratory approach. It is obvious that to do an in-depth study of this research methodology is not a scope of this thesis; however, discussing its characteristics is beneficial in order to gain a better understanding for this project.

Exploratory research is a research methodology normally used in the beginning of a study, when little has been done and when the researcher is more looking for confirmations of "emergent generalisations rather than an ensemble of a priori predictions" (Stebbins , 2008). Exploratory research is effective when a researcher faces a problem and he needs to find a possible cause for it. It is also useful to be familiarised with a certain field of study by, for instance, learning about the specific language used in this field. Lastly, we can also use this methodology to develop strategies that could be applied to tackle a given problem (University of Dayton, n.d.).

Exploratory research is described as inductive, opposed to a deductive (Stebbins , Exploratory Research, 2008). Indeed, whereas deduction is essential for a verification process, Stebbins explains that a deductive process cannot uncover new ideas on its own. He makes a comparison with the syllogism reasoning, explaining that if all A is B and all B is C, with a deductive reasoning, we can say that all A is C. However, such thought process will not say anything about any other proposition, D, E or F.

Most of the data outputted from exploratory studies is qualitative, although it is possible to use quantitative data (Stebbins , 2008). Since this field has not been studied much, in order to collect data, the most logical way is to produce it myself. To produce data, we can use different methods when following an exploratory approach. In exploratory marketing, the two main methods are focus groups and interviews (University of Dayton, n.d.). Whereas my research project is not about marketing, these methods perfectly fit my needs. I explained that this field has not been investigated much; nevertheless, it is situated on the verge of studied fields, such as OD and software design. This is the reason why it is beneficial for my research to do a literature study and find data related to my subject. Scrutinising this literature before creating my own data has the principle benefit of being relatively less time consuming, while providing me with data to start with and making my creation of data more efficient.

Now that I described the basics of exploratory research, it is necessary to explain how my thesis follows this framework. I described the state of the current researches on the matter of OD, and whereas the concept is not new to the academic field, I explained that there is little content in the specific field I aim to study. Moreover, as I presented in my research question, I do not state an already determined hypothesis. I seek to explore how OD influences the software development processes and find out if companies need to adapt their processes regarding this situation. I do not have any certain idea of how software processes could be modified, therefore I intend to explore the field, and observe how things really take place. Thus, I am not able to take an *a priori* stance to verify, because this is not the goal of exploratory research (Stebbins , 2008). Furthermore, my main goal is not to conclude with a precise set of instructions that APs should follow. It might turn out to be an outcome of my project, but due to the way this project is going to proceed, it is unlikely that I will be able to test and confirm such a hypothesis. This also supports the reason why the methodology of this thesis is an exploratory one.

Data collection

The collection of the data is an essential step of every project. A researcher needs to be able to raise questions relevant to his study, draw hypotheses and test their validity. In this section, I will firstly present the two methods applied to collect my data. Then, I will explain how I plan to use them to my work.

The first method I will present is the literature study, a secondary research method. Indeed, I aim to collect data from already accomplished researches, instead of creating data from research subjects or from launching new experiments (Housden & Crouch, 2003). Then, I will be able

to produce data on my own, through semi-structured interviews, which is the primary research method (or empirical data collection method) of my study that I will present afterwards.

Literature study

A literature study is the review of the existing literature related to a given subject of study. It is helpful to ground my explanation and get a first framework to work on.

We can identify several different kinds of source in a literature study, but it is essential for the researcher to work with content of the best quality possible. We can identify four main categories of articles (Olin & Uris Libraries, 2012):

- Scholarly;
- Substantive news;
- Popular;
- Sensational.

I consider to add another kind of source, to the four categories mentioned above and that would be the reports published by (or for) official organisations. Indeed, in my introduction, I referred a lot to reports ordered by the EU or other governmental structures, such as the MESPIR study. Whereas those reports are not peer-reviewed, I do believe we can consider them as trustable resources, being of the same level as scholarly ones.

In this literary review, I certainly seek to use the most reliable sources, which means that I will be mostly working with scholarly sources. The initial goal of this study is to provide an opportunity to draw a current picture of OD and to create a clear overview of this notion for the readers that are not familiar with the topic. This has been the base of my introduction. Secondly, through the selective literature research, I became able to frame my study, while observing how the topic has already been covered. However, the key part (and the most difficult part) of this study is to get insights into the impact of OD on the APs. As explained earlier, I could not find literature directly related to this topic. Nevertheless, it is still possible, to find material related to my frame of study, and I will present them in the results section.

Another benefit of this study is that it provides me with a better knowledge of the patterns necessary to discuss and analyse OD. Indeed, it is important to be able to draw hypotheses from the readings, but I believe that it is also necessary to employ the same vocabulary as the rest of the community uses, describing the same notions with the same words.

Semi-structured interviews

The semi-structured interview is a method to create data from interaction with people. They are different from structured interviews by giving to the interviewer the possibility to adapt his questions to the respondent's answers, by having a "flexible and fluid structure", when structured interviews require that all the questions are given to all the interviewees in the same way (Mason, 2004). This method obviously collects qualitative data. We can describe semi-structured interviews as a "construction site of knowledge" (Kvale, 1996). This means that contrary to other research methods, both the interviewee and the interviewer contribute to the creation of data. Therefore, the output of the interview is not only the answers of the interviewee, but the interaction between the two people.

This method has a certain number of benefits compared to other data collection methods. I demonstrated earlier that exploratory research usually uses qualitative data as its main data source. I am not looking to generalise the outcomes of my research, as I could through an extensive quantitative data collection process. Actually, through in-depth interviews, I will be able to produce more content than through a survey. Indeed, I do not have any a priori hypotheses, as explained earlier. Without discussions, I could not actually produce much data, since I would not have much to inquire. Through discussions, I will be able to reach the goal of my exploratory research: my focus is to get insights and to be able to better understand the field of OD and systems design. Semi-structured interviews are particularly suited for this situation (Gillman, 2000), or for the exploration of people's perception and opinion of complex issues (Barriball & While, 1994).

It is important for my study to run semi-structured and not structured interviews. A standardised approach would force the respondent to stay on my proposed framework of meaning, whereas with this choice I will let them to express more freely "their own perspectives, perceptions, experiences, understandings, interpretations, and interactions" (Mason, 2004). This is possible due to open-ended questions. Moreover, by doing structured interviews, I could miss some important points that the interviewee considers as essential ones and that I could have not expected.

Jennifer Mason identify several types of semi-structured interviews, where the methods used by the interviewer vary. The three distinct types are (Mason, 2004):

- *The ethnographic interview* is used in field research, where the interviewer wants to collect data mainly on the interviewees' interpretations and understandings;

- *The psychoanalytic interview* not only focus on the interviewees' conscious, but also try to explore their unconscious. To this matter, the interviewer for instance analyse the "interviewees' patterns of free association";
- *The life or oral history interview* is used to generate biographies, and can use specific documents such as photographs;

It is interesting to present those three different kinds because it might help the reader to understand how the definition of the semi-structured interview method is loose and can be adapted to satisfy specific needs. For this project, the type of interview I am going to use is definitely the ethnographic one. Indeed, what I am looking for is exactly what is described in its description, i.e. interpretations and understandings. We can also relate the type of interviews I will use to Kvale's qualitative research interview (Kvale, 1996), where the researcher wants to understand the subject's point of view.

Semi-structured interviews face several critics. The first one is related to the interviewer. As I explained earlier, the quality of the output is defined by the quality of the interaction between the interviewer and the interviewee. Respondents provide answers in different ways depending on the interviewer's sex, age, or ethnic origin. This is called the *interviewer effect* (Denscombe, 2007). Secondly, the way the interviewer leads the discussion influences how the interviewee answers. Like in any question/answer process, the interviewer must make sure to not push the interviewee's answers towards a specific direction (Newton, 2010). Moreover, there is always a risk that the interviewee chooses to answer in a way he thinks is the best for the situation, even if it is does not necessarily reflect his opinion (Gomm, 2004). The last big criticism raised upon semi-structured interviews is the impossibility to make a comparison between several interviews. Since questions can change from one interview to another one and the focused topics can be different, we lose this possibility (Mason, 2004; Newton, 2010).

However, most of these issues can be solved in general or in my more specific context. Firstly, whereas the interviewer effect can indeed be a problem, it is not necessarily the case here. Indeed, the different characteristics of the interviewer usually influence the "amount of information people are willing to divulge and their honesty about what they reveal" (Denscombe, 2007). Regarding my study, it is unlikely that my gender, ethnic origin or age would be a problem for the respondents, especially since the topic I am going to cover is not a sensitive one. The other weaknesses related to the interviewer can be overcome with an adequate preparation. It is obviously possible that I might make some mistakes at this level, but it is not something to consider at this point, but to carefully keep it in mind while analysing the results.

The last problem is about the coherence of interviews, which could not be easily compared. Whereas it is true, it is possible to raise two counter-arguments. Firstly, some degree of comparison might still be possible between interviews since semi-structured interviews do have some sort of structure that the interviewer follows (Newton, 2010). Secondly, we must remember that we are not looking at comparisons between interviews. Such things can be interesting, but we are more interested in patterns the respondents' answers. Indeed, this is a qualitative research and not a quantitative one , thus the sample size does not matter in the same way, as long as we do not want to generalise our results to the population at large (Englander, 2012). We can also argue that this drawback has its benefits: the user, by being given a certain freedom by the interviewer, can build its own coherence within his answers (Newton, 2010).

A semi-structured interview has several strengths compared to other methods. I previously presented the creation of an interaction between the two people. A benefit of this situation is that it allows the creation of "intimate, trusting and empathetic relationships" where "respondents feel able to disclose the truth" (Gomm, 2004). Another point useful to my case is the fact that this method gives me an opportunity to build a relationship with the interviewees on a long-term basis (Barriball & While, 1994), and grants me the ability to get feedback later in my project. It might be also be beneficial that semi-structured interviews tend to lower the difficulties of working with non-native speakers, which could have occurred in my research (Barriball & While, 1994). The last important point is the possibility of using probes (Hutchinson & Skodol-Wilson , 1992). Probing is a method applied during interviews to allow to "acquire additional information (Mahoney, 2004). They can be simple questions like "Such as?" or "Could you describe that process a bit more?" (Mahoney, 2004). Interviewers can use these additional questions when they believe that the interviewee might have something else to say about a topic. When an interviewee finishes answering a question, using a probe is a way to ensure that there is nothing else that was forgotten.

Probing ensures a certain reliability of the data by (Barriball & While, 1994):

- Allowing the "clarification of interesting and relevant issues raised by the respondents" (Hutchinson & Skodol-Wilson , 1992);
- Providing "opportunities to explore sensitive issues" (Nay-Brock, 1984; Treece & Treece, 1986);
- Helping the "respondents recall information for questions involving memory" (Smith, 1992).

Being more precise, by probing, we can significantly increase the quality of the answers.

Use of the methods

The first step of the research is to gather a sufficient amount of knowledge about the OD topic. A part of this knowledge has been disclosed during the introduction, where I presented the key notions of the topic. Through the literature study that I will present in my next chapter, I hope to raise a certain amount of questions, which will need to be investigated in order to find an answer to the research questions. Those first questions should provide me with sufficient material to prepare the semi-structured interviews. Since there is very little academic content on the perspective of this project about OD (i.e. the Applications Providers), those semi-structured interviews should give me the opportunity to describe the process of building applications on top of OD, we could say that for this part, the interviews have a role of case studies. These interviews will also offer me a possibility to discuss my findings with the concerned actors and get valuable opinions. Hopefully, these discussions should raise other questions, in a process of cascading interviews: the first ones would be more exploratory, where the last ones should deliver specific and valuable answers.

We can visualise both the literature study and the interviews as parallel processes, where the data collected from these sources helps to gather higher quality data. Indeed, the literature is the initial starting point, letting me design my first interview, giving me a sufficient knowledge of the subject of OD, and helping me to find themes that need to be discussed. However, the interviews will raise new questions and will give me new elements to work on. At this point, I will be able to go back to the literature. It is obvious that a literature study cannot be exhaustive. Some points are likely to be excluded by the researcher (who cannot think about all the possible keywords or notions related to the studies topic), and others can be regarded as less important. The content produced through the interview and its analysis can direct me toward sections of the literature I would not have studied yet. This means that I will need to have a certain amount of new data before starting the second interview. Therefore, I am able to adapt my interview schedule to discuss those new points.

Consequently, it is possible to see my data collection process as an iterative one, where at each step I collect some new data, and I can modify the subsequent interviews. Since my approach is an exploratory one, I believe that it is interesting to make a comparison with concatenated exploration. Concatenated exploration is a method theorised by Robert A. Stebbins, which is used to explore a field by subsequent studies. The term describes "a longitudinal research process and the resulting set of field studies that are linked together" (Stebbins, 2006). I cannot

considerer my methodology as being concatenated exploration since it describes a set of studies (Stebbins gives the example of "Robert G. Burgess [who] spent over twenty years conducting a variety of interrelated studies"). However, the ideas behind this method are actually close to mine. Firstly, he describes how we can relate successive field studies to each other, and it is possible to use his findings to describe the relationship between my different iterations (literature study + interview). He explains that the methodology is different from a chain of studies, since not all links have the same importance, and the studies are "not only linked but also predicated on one another". He also describes that like planned in my methodology, "later studies are guided … by what found in earlier research". An important point that I will have to consider is, as he conveys in his study, that any findings need not to be "treated as given"; thus, at the later steps of the project, I will have to be on "the lookout for disconfirming findings from the present study". Moreover, he states that looking to disconfirm these findings has the benefit of ensuring their validity, which I will also consider for my study.

This methodological approach can be described by a model (Figure 1 Methodological approach). There, we can see how the research is done through iterations, backs and forths from the literature to the interviews. Each interview might guide me towards new or more specific literature, and each consecutive interview should be improved based on the data collected so far. Finally, the literature also helps to analyse the collected data. Indeed, in the first part of my result sections, I will present themes coming from the literature to confront the interviewees with questions related to these themes. Moreover, the same literature might actually give some answers to these themes.

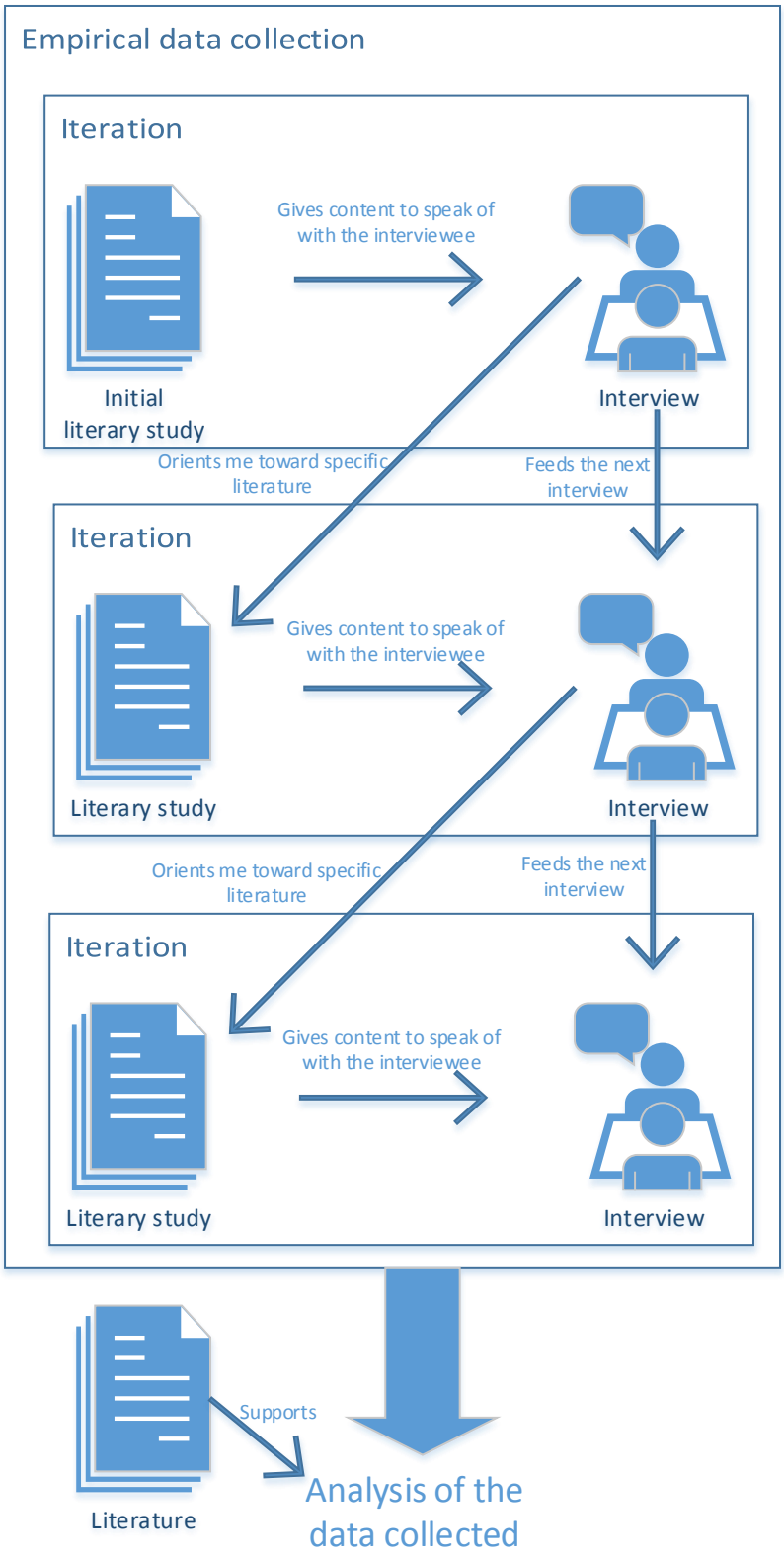

Figure 1 Methodological approach

Planning of the interviews

In this section, I will present the interviewees and will explain my selection process. Indeed, selecting the appropriate respondents is an essential step of the data gathering process, and the results will heavily depend on the quality of this selection (Englander, 2012).

The goal of my interviews is to collect empirical data about creating software with OD. To reach this goal, to interview people who had this opportunity is the logical step.

My target is two main categories of people working with OD. The first ones work in a professional setting, whereas the second ones can be considered as working in a volunteer setting. Professionals include people working in companies using OD with a business-oriented goal. They might have other motivations such as ethics, or they might believe in the benefits for the society of OD (such thoughts exist within the Open Source community), but their main motivation is business. The second settings come to people whose prime motivation is not profits. They might be individuals joining a hackathon or working on "side-projects" (projects developed on their free time, often without expecting any other benefits than self-promotion). They might have several different motivations. It can be a belief in citizen-driven initiative and that contributing to this movement is beneficial for the society. They might also want to experiment with this new kind of data, expecting to gather knowledge about it, which might be useful later. These motivations are not a source of interest here and I will not give more details on those. Firstly, what I described is mostly based on assumptions and the profiles of the people that I will present in the Table 1 Interviewees. Secondly, whereas the motivations of these people is not the prime goal of my study, I will have a look at it through the interviews. However, it is interesting to point at these different profiles, since the methods of these respondents might differ, from a strict professional context to a more "in the wild" one, where developers do not necessarily have to follow the same methodological constraints.

I have consequently selected four people to interview, with different profiles, presented in the Table 1:

Participant	**Overview**	**Setting**	**Interview**
Software developer	Used OD during his work	Professional	1st
Software engineer	Participated in two hackathons	Volunteer	2nd
Software engineer	Participated in two hackathons	Volunteer	3rd

Project manager	Supervised research projects using OD	Professional	4th

Table 1 Interviewees

I explained earlier that my project is an explorative process. Moreover, I explained in this chapter how the data collection is an iterative process, each interview fuelling the next one. Therefore, in order to maximise the quality of the output, it is crucial to plan these interviews in the right order, since each interview guide is an improved version of the previous one. The last respondent, the project manager, is the person with the biggest and broadest knowledge regarding the topic. Therefore, I plan to interview him the last, the three other interviews will be used to gather knowledge to make this last interview as efficient as possible. The two engineers have a specific knowledge on OD but it is not clear that they used it apart from these special occasions. Starting with them would probably not be the most satisfactory choice, since we would have a limited number of themes in common (i.e. subject he can speak about, and that I am aware of). This is the reason why I choose to interview them after the software developer, who may have a good general knowledge of the subject, and that would be helpful to extend my insights on the field.

In brief, firstly I will interview the software developer, in order to treat general questions and have a first overview of the topic. Then I will interview (separately of course) the two engineers to increase my knowledge in the specific fields of theirs. Finally, I will question the project manager, with whom I will be able to have the most in-depth discussion.

Content analysis

Once empirical data is collected through the interviews, I am going to analyse it, i.e. make it usable. This analysis will be carried out by applying content analysis and following the methodology presented by Denise O'Neil Green in the *SAGE Encyclopaedia of Qualitative Research Methods*. The aim of this analysis is to describe the content and to characterise its sense. Most of the definitions emphasise the systematic nature of this analysis (Kaplan, 1943) (Holsti, 1969), or also the matter of objectivity (Berelson, 1954). In brief, this process consists of an exhaustive and systematic description of the content, and a treatment to make it usable. By being systematic and aiming for as little distortion of the meaning of the data as possible, the output evolves from the data, covering all of its content, without altering what the interviewees meant.

There are several approaches to content analysis (Franzosi, 2004), but I will select a common one, where the main idea is to build a global understanding through an iterative process, creating

codes from the raw data, then building categories of these codes, and finally themes, which are the highest level of understanding. Here, I used the process and perspectives on coding described by Corbin and Strauss' work with grounded theory (1990), O'Neil Green (2008) and Lockyer (2004).

It is not possible to analyse raw data. In order to manage to get an understanding of a set of interviews, it is necessary to break down the raw text into analysable units, to get the "concepts" out of the text, which are the "basic units of analysis" (Corbin & Strauss, 1990). The action of building these manageable units is also called coding, which is "a systematic way in which to condense extensive data sets into smaller analysable units" (Lockyer, 2004). A text is in itself complicated to work with, providing limited ways to organise or categorise its data, theories. The researcher needs to work with concepts representing the data: "incidents, events, and happenings are taken as, or analysed as, potential indicators of phenomena, which are thereby given conceptual labels" (Corbin & Strauss, 1990).

After coding the data, the researcher is able to create categories to "conceptually organize findings" (O'Neil Green, 2008), grouping concepts that "pertain to the same phenomenon" (Corbin & Strauss, 1990) into categories. Such definition also encompass themes, which we can also see as higher-level categories, being a macro-level of analysis, compared to the meso level that categories are (O'Neil Green, 2008). For instance, in a set of interviews, we could see that an interviewee speaks about his interest in solar energy and another one expresses his about wind energy. These would fall into a category of "renewable energy". If they also speak of bike tracks and public transports (which would fall in a category of "non-polluting transportations"), we could study the theme of "ecology".

In this thesis, I will develop categories inductively (opposed to deductively), without a predetermined list of categories (O'Neil Green, 2008), which is consistent to my exploratory process. However, it is possible to go back and forth for both analyses, also building categories from previous studies (such as my literature study).

Now that I specified the way I had learnt from the data, it is important to explain the organisation behind that work. Firstly, I convert the audio of the interview into a manageable text log. This is not a full transcript, but a detailed list of things said during the interviews. Despite the fact that I conducted the interviews in French, this transcript is in English. Moreover, I tried to stick with their wording as much as possible. Here, there is little interpretation of the content, the logs are supposed to simply represent the data under a written

format. The second step is to make the data to be usable more easily and therefore to write a summary of the content of all the interviews. This summary simply follows the structure of the interview guides and for each point, it is mentioned which interviewee agreed or disagreed with that specific point. Briefly, it is a friendly readable version of the content of the interviews. It is possible to find both of these documents in the appendix of this thesis, since they are not properly part of the analysis (the four logs are the appendix F, G, H and I, the data summary is at the appendix J)

Here the coding starts, based on the summary I just mentioned, followed by the content analysis described previously.

Results

In this section, I will present the results of my data collection process. I will use two kinds of data sources: the literature, and the interviews. These methods are not linear (cf. Figure 1 Methodological approach). Because of that, I cannot present the data from the literature study, and then from the interviews, however, the literature study can be broken in several pieces and then interrupted by the interviews.

Literature study

As I explained in my methodology section, the literature study was not done at once, but was compiled of several steps, the interviews directing me towards specific literature, which consequently helped me to improve the quality of the interviews.

My first objective of this literature study was to be familiarised well with the topic of OD (which I presented in the introduction). With application of this knowledge, I was able to identify emerging topics of discussion, which I used to build the first interview guide.

Before presenting the topics that emerged from the literature, it is essential to explain what their precise role is. To identify the emerging topic of discussion, I am not exclusively looking for high quality topics of discussion. It is not noteworthy if I believe that one topic is more likely to deliver an interesting discussion when comparing it to another one. Whereas I work with qualitative data in most of my thesis, in this part I am looking to produce content in a quantitative way. Indeed, I proceed in an exploratory way, therefore, my goal is to discuss as many topics as possible, the data will anyway be filtered later in the analysis (if a topic does not yield any interesting data, it will not simply be taken into account).

Consequently, it is possible to use different kind of sources for these topics, more or less reliable, what matters is to find as many sources for these topics:

- Some of the topics I discuss are mentioned in only one paper;
- Some other do not come from academic literature;
- Some topics will come from my personal experience of working with OD. Using the personal opinion of the researcher is usually not a recommended method when generating data, but I need to keep in mind that this opinion is not in itself data I will

analyse later. If the interviewees do not agree with this topic, or do not have anything to say about it, later it will not be used, and thus will be discarded after the interviews.

After a table showing the source of each topic, I will present them. I will give a brief overview of the topics, but I will not give an interpretation of them. It is essential to get familiarised with them in order to be able to run a discussion with the interviewees about those topics, but as explained, the goal of this section is to produce content usable in the interviews, and not to make an analysis of these hypotheses which will be made later in this chapter.

Theme	Source
Open Data as a trigger to innovation	Academic
The difference between internal and external data	Academic
Open Data platforms	Academic
Linking datasets	Academic
Changes to the organisation	Academic
Technical constraints of software development	Personal experience
Standards and format of data	Personal experience

Table 2 Summary of the topics of discussion

Open Data as a trigger to innovation

Some of my findings highlighted a link between OD and innovation. For instance, Maria Teresa Borzacchiello and Max Craglia discuss how freely accessible geospatial information could increase the production of innovation (Borzacchiello & Craglia, 2012).

We can also see OD as one of the elements facilitating user-driven innovation. Eric von Hippel describes how it became easier for users to innovate by themselves, having a cheap access to resources, tools, software or hardware (von Hippel, 2006). We can consider that OD is one of these resources and its existence therefore facilitate the innovation process. It is also possible to look at the situation from the perspective of a manufacturer (i.e. organisation producing goods or services in order to sell them). There are cases of companies (General Electric, StataCorp for instance) giving to their users the possibility to customise the products, to adapt them to their needs. Consequently, the companies pick the best user innovations and bring them back to their

product (von Hippel, 2006). Application providers could consider a similar process with the opening of their data.

The difference between internal and external data

When discussing with companies about the use of OD in software business, Yulia Tammisto and Juho Lindman interestingly discovered that companies do not necessarily relate *Open Data* to its proper definition (i.e. the *Open Definition*). They put on the same level "internally open data" and "externally open data", where the first category is actually about data produced within the company and only accessible to its employees and the second category relates to what we define as OD (Tammisto & Lindman, 2012).

Open Data platforms

Researchers wrote a lot about the OD platforms since they are the principal access point to open datasets. There are problems with the quality of the datasets on the platforms and with the ways to improve them (Eberius, Braunschweig, Thiele, & Lehner, 2012). It would also be interesting to discuss the shift to more advanced "web 2.0" platforms, where users can be empowered and become "pro-sumers" of data (Charalampos, Loukis, Charalabidis, & Zuiderwijk, 2013).

Linking datasets

After finding a dataset, to identify related datasets is actually often a challenge. Indeed, in order to succeed, these datasets are required to contain information that makes easier for the user to find other datasets. Two of the main techniques to accomplish this goal is the use of meta-data (Zuiderwijk, Jeffery, & Janssen, 2012), or of Linked Data (Bernes-Lee, 2006). Both of these methods have been explained in the introduction of this thesis.

Changes to the organisation

People in organisations working with OD seem to believe that it makes an impact on the organisation of the AP, improving communication and decision-making processes for instance, or helping to develop new services (Tammisto & Lindman, 2012).

Technical constraints of software development

There is much literature on the evolutions of the software development methodologies, from the shift to agile methodologies between the nineties and the beginning of the 21st century, to the use of FOSS by companies. The first shift was caused by the discussion between developers who figured that classic methods were not efficient when developing software, whereas the

second comes from the arrival of new tools accessible to companies. Does developing with OD sources require different techniques than in a more common process? For instance, it is possible to discuss the reliability of these sources, when there is no organisation taking responsibility of the quality of the files.

This topic is something I had the opportunity to work on and analyse during my 9th semester project, where I accomplished an internship in Orange Labs (R&D department of Orange), where I discussed the use of agile methodologies in research projects. Whereas OD was not the focus of my paper, I had the opportunity to work with this type of data, and therefore considered that this topic should be investigated more with the interviewees.

Standards and formats of data

Similarly to the previous topic, my working experience with standards and formats of data comes from my experience gained during my last internship, however, I did not discussed this point during my project. One critical point when working with data is to be able to work with the standard used. Some standards are rather well formalised at an international level, such as the metric system (but the USA, for example, do not always use it). When you start looking at more complicated systems, such as geographic coordinate systems, things become trickier. There are several standards when looking at Earth as a whole (WGS 84, GRS 80), and numerous national systems: the wiki of OpenStreetMap shows the differences between the Austrian, the Spanish, the British, the Swedish, the Dutch and the French (which is in itself divided in eight sub-systems) systems.

However, those standards refer to common objects (despite there are different ways to describe geographic coordinates, they all refer to the same thing). This means that there is a limited number of ways to describe the objects, and that there are ways (not always simple ones) to convert the description of the object from one standard to another one. However, many things have not been standardised. Let us imagine a fictive case, in order to see how things are complicated. If today the US federal administration decides that all the administrations need to publicly share what are their furniture. Two different administrations could share these documents:

Administration 1		Administration 2	
Type	Desk	Origin	USA

Made In	China	Main material	Wood
Material	Wood/Metal	Secondary material	Plastic
Use	Reception	Category	Desk
Price	$500	Price	500

Table 3 Example of non-standardised datasets

This is a simple example. If an AP wants to build statistics on the furniture, it could manage to do it by matching the different cells (type and category refer to the same thing here). However, the problems start when a company wants to do it programmatically. There are around 3.000 counties in the USA. If they do not present their similar data in a similar way (i.e. with the same format), it gets more complicated to work with it. In an interview with the *Computerworld*, the chief economist of Zillow (a company working with OD, valued at more than 1 billion dollars) describes the problem when working with public estate data and says: "Anyone who has worked with public record data knows that real estate data is among the noisiest you can get. It's a train wreck". These are the consequences of a lack of standard formatting (Waxer, 2013).

Construction of the subsequent interview guides

As explained earlier, the content analysis method requires that I treat all the interviews once done with all of them, in order to read patterns within the different interviews. However, due to my iterative data collection process, it is also necessary after each interview to draw an overview of the subjects discussed and see if anything should be added to the interview guide of the next interview. The summary of the three first interviews can be found in the appendix E (there is no summary of the fourth one, since this was the last interview). In this section, I will present the evolution of the interview guides after each of the three first interviews, i.e. I will give some details on the new topics raised by each interview and what they bring to the next one. This evolution is then summarised in the Figure 2 Development of the interview guide.

First interview

This interview brought two new topics to discuss in the second interview guide. The first one is named "understanding the data". Here, understanding covers the questions of assessing what the data covers, what it does not, to look at its value, what can be done with it, its meaning, etc. Indeed, the usages of data are not obvious, and a thought process might be required to work with it. Secondly, I added a technical theme regarding the creation of API between the data and the developers, i.e. a layer permitting the developers to abstract the data. Lastly, I added a

question to the "changes to the organisation" theme about the fact that open internal data might empower the employees of an organisation.

Second interview

A reflection on the process of understanding data made me think about the more global process of working with and designing data, and the biggest part of the question is if the idea comes from the data, or if the data is selected after finding an idea. This brings in the interview guide a new topic for discussion, entitled as "development process". We also talked about the role of the journalist in understanding the data, which led me to add this as a question to the topic of "understanding the data".

Third interview

This third interview did not reveal anything much specifically new that needs to be brought to the next interview guide. Nevertheless, it increased my knowledge on the topic, which was obviously useful for the discussion with the final interviewee.

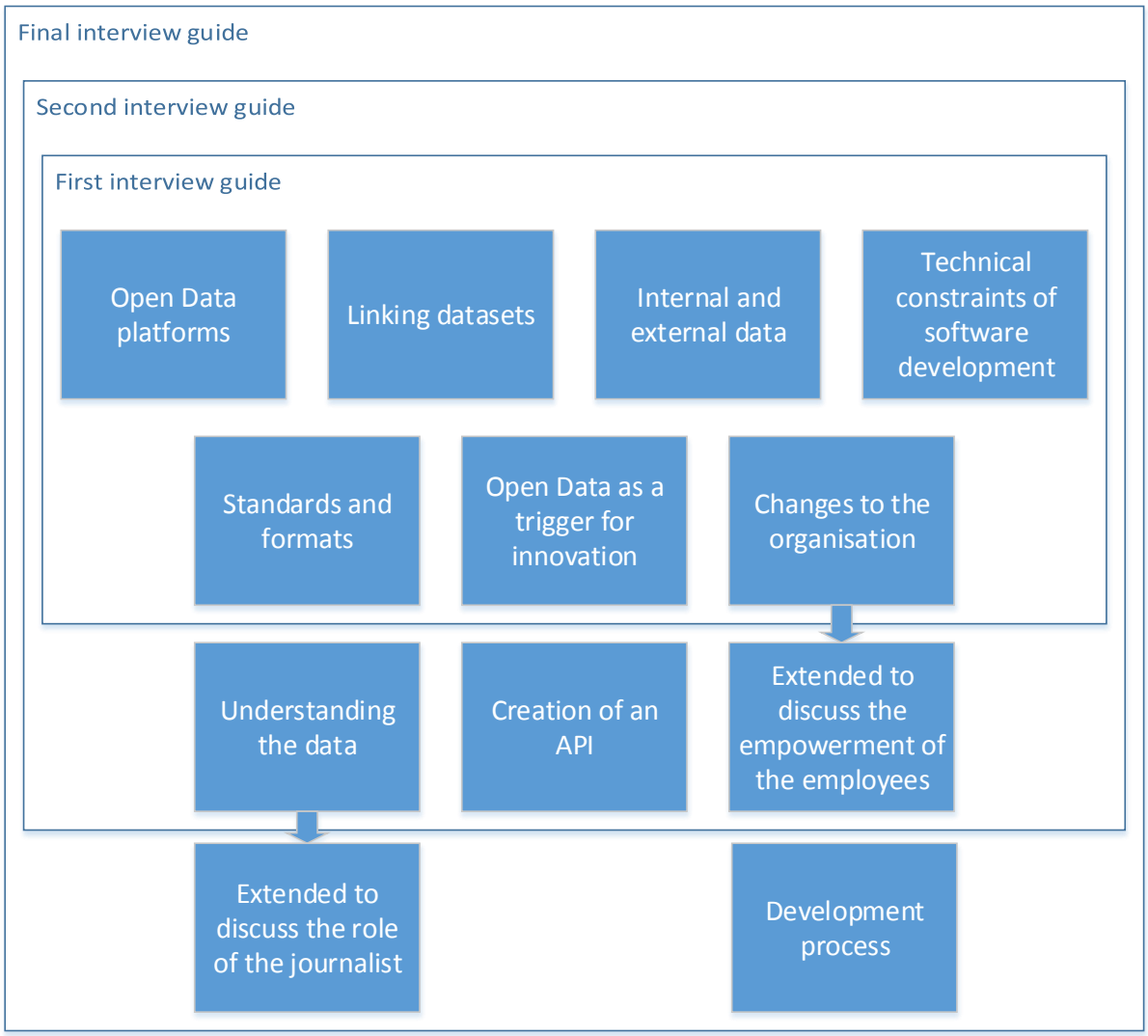

Figure 2 Development of the interview guide

Content analysis

I explained in the methodology section the process of content analysis, where I code the data to render it analysable, then create categories out of these codes and finally draw the themes. The summary of this data can be found in the appendix E, which contains the codes. In this summary, I highlighted the parts of the text related to each other.

Categories

In the data summary (Appendix E), I proposed a first level of analysis with the creation of ten different categories, covering all the relevant data produced by the interviews. In this section, I will present these categories and their meaning, referring to the interviews as much as possible.

Reliability of the data

When working with commercial data, the company providing it usually takes responsibility for the accuracy of data, and guarantees the absence of possible mistakes in it. However, when it

comes to OD, such responsibility is usually not engaged, the publishers are fine to provide the public with data, but do not with support for it. Moreover, whereas crowdsourced data can be considered as reliable (for instance, according to Haklay, crowdsourced geographical information "can reach very good spatial data quality" (Haklay, 2010)), there is still the issue of no one being responsible for it, which, in a corporate environment, can be a problem.

The last interviewee discussed about how the nature of the data used might also depend on the reliability needs. He explains it through an example:

> *"Imagine OpenRunning, I use their data when I am training, because I am fine with this data, and it is free. When I start to run a bit more seriously, and I start to join the community of marathon runners, I use the marathon data that I pay 2 euros a month."*

The interviewee means that reliability of the data matters if the project is critical. As long as it possesses few mistakes in the datasets, it is not problematic for everybody; therefore, the reliability of the data is not necessarily always a concern.

Preliminary work on the data

The interviewees discussed about the matter that a certain quantity of work needs to be done on the data, in order to work with it in a project, or to increase the efficiency of the team. Firstly, the first interviewee emphasised the fact that OD is not a service: it is raw data, just published as it naturally is. The developers need to do everything by themselves. This includes several categories of steps to do.

Before starting to work with OD, the organisation has to decide which datasets to use. Indeed, sometimes OD is the only available data, but it is often not the only existing one. There might be commercial organisations proposing their own data (for instance, both Google Maps and OpenStreetMap propose geographical data), or companies proposing their services on top of OD (Data Publica, or OpenDataSoft). Consequently, when deciding which offer to choose, the AP needs to filter and assess them.

We can see the first step when working with the OD as a basic one, including cleaning, correcting, updating the data and eventually converting the files' formats. Indeed, the interviewees stated that the quality of most of the files found on the OD platforms is relatively low.

Furthermore, as explained in the previous category, sometimes, the certification of the data is mandatory, which would make a certification process necessary at some point.

Lastly, it is important that the data suit the company's specific needs. First of all this includes a compatibility with the company's technology (putting the data in a database, create an API, etc.), but eventually also include the creation of a piece of software that the third interviewee called a "wrapper":

> *"I will give a personal example. Consider the timetables of students in a school, if you convert them to an ICS format, students are, I hope, on time more often […], I put the data in a usable way […] people remixed it, remade it"*

Here he gave the example about the time when he was a student and he scrapped the data of the school's portal for the timetables, and converted them to an open standards. It allowed the students to put their timetable on Google Calendar, and get SMS reminders before each class. Here, the data was already accessible, however, its format, the way it was proposed was complicated to use. The piece of code he used to scrap the data from the portal was consequently giving value to the data.

Skills

The last three interviewees discussed about the various sets of skills that are useful for a company working with OD. The first one is the role of "data specialists" (the double quotes indicate that the term is not mine, but that I am using the one given by the respondents). These people have technical skills when it comes to work with data. They are able to process it, parse it, change formats, convert it, etc. The second category of technical skills necessary when working with OD (and especially with large amounts of OD), is the skills possessed by a "data scientist". A data scientist relies on mathematical and statistical methods to show patterns in data, build insights and mathematical models of it. The last technical job needed was described by the third interviewee as the "XML expert". This kind of person is able to make complex models of data. Indeed, two-dimensional tables cannot represent everything.

> *We need people able to "make models of the legal structure between companies, of the correlations between probabilities of consumer defaults, etc."*

There are other categories of people that are interesting for a company working with OD. The first one could be described as a "journalist". A journalist needs to look for information, understand its meaning, figure why it is interesting, and then, present it in a readable and efficient way. Whereas the third interviewee believes that searching for data is not much of an issue (much of the interesting content is already known or is easy to find), the second, third and fourth interviewees believe that this process is important, and especially the one related to the understanding of the data (which is covered by another category, discussed later).

The last person described during the interviews is the "designer/artist". This person is by definition able to think differently than an engineer or a manager and can help to produce something "out of the box". Indeed, the fourth interviewee believes that having a creative background, or being trained to design techniques (for instance, he mentioned design thinking) helps to generate ideas that we can describe as non-conventional, or innovative.

Value of the data

We can discuss the value of OD at several levels. OD has a value for the AP, by potentially reducing its costs (time and money), since the information is free and do not have to be created. As discussed earlier, the value of the data increases when shaped (or wrapped) in a correct way. This implies that an AP can make a capital gain by:

- Reducing its costs by using OD;
- Increasing the value of owned data through wrappers and reshaping.

On the other side, OD can have a negative effect on the value of data owned by a company. Indeed, precise and exhaustive data's value will be lowered by the existence of cheaper/free data, even if the latter is not as precise or correct. As explained when discussing the reliability of data users who could pay for good data might rather not spend anything and use OD. This implies that the owner of the more valuable data will have less opportunity of selling his data; consequently, its value will be lowered. The third interviewee even predicted that OD will mean the end of organisations having a monopoly on some kind of data (for instance, map providers). Thus, as he explained, the value will not be in the data itself, but in the service which is on top of this type of data (a wrapper being a low level service in itself).

Design process

When we discussed the different steps of designing a service/application, the first two interviewees stated that data is both a catalyst that permits the realisation of ideas and a mean to get new ideas. However, the third and fourth interviewees disagreed on the second part of

that statement. Indeed, the third one said that whereas you can have an idea with OD, of how to display your skills or your technology, you get an idea of a useful service when you identify and try to solve a need, not by starting from the data. The fourth interviewee's opinion was not that strong, but he claimed that through his years of experience, he always observed that there is a process of "first idea then data" not "first data then idea".

> *"I do not think there are people who say 'this dataset is cool, I will use it'. … There are graphic designers who will do some demos … but it is a 'happy-few' [in English during the interview] thing, for demonstrations, proofs of concept."*

Talking about methodology, the fourth interviewee argued that agile methodologies when working with OD are important, and especially when doing research projects (i.e. when the outcome of the project is not strictly defined). One of the characteristics of OD is that you cannot decide how you will get the data, what its format will be, when it will be updated, how detailed it will be, etc. This means that you do not have any form of control on the data. Consequently, working through iterations, being as close to the data as possible is important, in order to be able to change the process if needed. Staying close to the data means that when developing an application, the developers should use the data directly, and work with it. When developing a complex software, it is a common practice to split the project into several parts, and in order to test the developed parts, to abstract the other ones. For instance, let us consider an application that creates visualisations of data. We can divide such an application into two parts, a database, and an interface showing the visualisations. From a technical perspective, it can be tempting for the developers to avoid working with the data from the beginning, using for instance files to simulate the database, in order to focus on the visualisation. However, the fourth interviewee thinks that is better to use the real data from the beginning.

The interviewee summarised his opinion about this subject by arguing for flexibility during a project:

> *"You have a starting point, then an ambition, and eventually an objective you would like to reach. There you need to be ready to adapt your work… there are three points: what you hope to make, the technic to do it and the data … . You have to juggle with these three points, not to hesitate to modify one or the other to go further."*

Characteristics of the organisation

When discussing how an organisation should be organised to be able to work efficiently with OD, two main points were given. The first one is the efficiency of a flexible organisation, when communication is easy. Indeed, OD can come from multiple sources, and their discovery is not always obvious. Consequently, in a pyramidal structure, someone who is at the bottom and finds data would not necessarily be able to exploit this data easily. The second point is closer to the ways that teams work: the fourth interviewee considers that multidisciplinary teams and the use of techniques like design thinking are beneficial for organisations.

Description of the data

When working with data, to be able to describe it is usually one of the things to do. Indeed, a description is often a first step to the understanding of data (this will be explained later). On a different level, a description of a data is powerful when this description is unique: having unique identifiers for objects permit us to refer to them in an easy and unambiguous way. When it comes to database this can take the form of primary keys, but in some businesses, it can be also expressed with some specific conventions. The third interviewee works in a bank, and he described how the usage of naming rules (and institutions) makes transactions possible between different institutions in different countries.

> *"The SWIFT messages […], or the BIC, or the IBAN […], the emergence of the symbologies or numbering systems, open and usable by everyone, without any licence, the fact of being able to communicate through one medium, brought value. The phone numbers, when it became international, and easily interoperable, created value. When the SEPA (Single European Payment Area) started, it brought a creation of value. An open symbology brings a creation of value."*

External opening of the data

During the interviews, we made a list of positive and negative consequences of opening data to external organisations. Opening data can foster external innovation. By external innovation, I mean services or products based on the data that the company could have eventually made by itself. For instance, there are plenty of web applications using Twitter's tweets, to provide users with services, to do special search, etc. For instance, I use the website IFTTT.com to monitor my Twitter feed. If I favourite a tweet which contains a link, it will be bookmarked, so I will be able to read the webpage later. Such action would not be possible without Twitter's public API (or would be technically complicated).

Twitter benefits from this situation. Firstly, its users' experience is increased. Secondly, if Twitter finds an application especially interesting, it can copy it or buy it (for instance, Twitter bought TweetDeck, an alternative client for 40 million dollars in 2011 (Le Monde, 2011)). Lastly, since the accesses to the data occur through its API, Twitter never loses control on the data: if usage of their data displeases them, the authorisations can be revoked.

Another important benefit of opening data is its impact on the company's image. Releasing data publicly can be a way to promote the company's data or to show a will to transparency. Lastly, it can also offer strategic benefits, through the creation of a community around the data, or advantages against competitors (by lowering the value of their data and making the company's data more interesting).

On the other side, such opening can have negative consequences, which can be predicted and avoided by carefully planning the opening the data. In case of data monopoly (or data with a higher quality than competitors), the AP would lose a competitive advantage. Moreover, the company needs to pay attention to the privacy behind the data. Lastly, opening data forces some sort of transparency, which is not always what a company is looking for.

Internal opening of the data

Making data easily accessible within a company, for instance between its different departments offers several benefits. The first and most obvious one is that it fosters innovation by facilitating the reuse of data. The second effect is more a secondary benefit, since it could boost the internal communication of a company and improve its internal culture.

> *"It is like Open Data in the civil society, it could make the management flatter: all citizens have access to political data, like all employees have access to the company's data."*

Understanding of the data

The last category revealed by these interviews is related to the understanding of data. This notion emerged when we were discussing the design process. When looking to usage of data, it is important to understand the possible ways to use it, its value, its limitations, etc. This is part of the work the "journalist" does. Moreover, this has been mentioned as one of the differences between external and internal data. Indeed, with internal data, it can be possible to get access to a support, contacts to discuss the data, and to know more about it.

Themes

After building a second level of understanding, namely the categories, I grouped them to create themes. In this section, I will present four themes, after a table recapitulating the correspondences between categories and themes.

I grouped the data, which gathered through the interviews, in ten categories. This categorisation was a mandatory step, in order to be able to navigate easily through the data. However, ten independent pieces of knowledge is a too high number to lead a constructive discussion. This is why we will try to understand how these categories are related to each other, i.e. what themes they do belong to.

Category	Theme
Reliability of the data	Working with the data
Preliminary work on the data	Working with the data
Skills	Preparing the organisation
Value of the data	The data
Design process	Working with the data
Characteristics of the organisation	Preparing the organisation
Description of the data	The data
External opening of the data	Open innovation
Internal opening of the data	Open innovation
Understanding of the data	The data

Table 4 Categories and Themes

Working with the data

When reading through the different categories, the first common point is that OD can imply direct and immediate changes to the way APs work with data. Indeed, when working with OD, it is essential to consider the reliability of the data, to do a certain number of preliminary steps when working with it, and finally, adjust the design process. It is possible for APs to adapt the

way they work with data by following these advices. They will not have to change model of their organisation, but simply to adapt its methods.

Preparing the organisation

Whereas the changes discussed in the last theme do not require changes in the organisation, but simply of its methods, this theme covers what I call structural changes, i.e. changes, which would require to significantly moving the structure of the organisation. Such changes would have to be planned before working with OD, and cannot be introduced on the fly.

Changing the characteristics of the organisation, by changing its culture, improving its internal communication, introducing techniques such as design thinking, or gathering multidisciplinary teams needs to be considered before a project that uses OD.

The data

Knowing more about the data, understanding it, determining its value, finding the relevant datasets is also a theme that needs to be discussed. What these categories have in common is the fact that analysing the data is beneficial to the AP. They do not focus on how to work when already possessing the data, but what data an AP should look for, how it could determine what the data is worth. Moreover, this theme affects the other themes For instance a deeper understanding of the data might let a developer write a better wrapper for the data.

Open innovation

By giving access to the data to more people, an AP can provide them with tools to innovate on top of this data. People with various profiles and needs would then be able to use whatever they need in order to come up with ideas of usage for the data, or concepts of how to use these data to solve problems they have encountered.

Discussion

In this section, I will discuss my findings, which were presented in the previous chapter. As explained in the Figure 1 Methodological approach), I will relate them with the existing literature on the topic. I will also discuss how they are able to answer my research question, while showing the limitations of my findings.

Proposition for a design process

I will now propose a model of the global design process, as described by the interviewees. This design process is not only composed of a set of tasks that I will describe, but also of skills that an AP should have. Moreover, this process should be agile.

An OD design project starts from an identified need. Then, the AP has to come up with a solution, an idea, how to solve this need. In *Democratizing Innovation*, Eric von Hippel discussed how a lead user is more likely to identify a need and come up with an idea than a manufacturer, expecting to "gain relatively high benefits from a solution to the needs they have encountered there" (von Hippel, 2006). Finding ways to identify needs also benefits from both internal and external data opening. I explained when describing these two categories how they can foster innovation. While external opening do not always imply internal innovation, I explained that this innovation can still be brought back in the AP, through copying or buying. An easy access to internal data also gives means to innovation. Steve Yegee, former Amazon's employee explained how their decision to open all their data through API was a great resolution, and how it let them to transform the "infrastructure they'd built for selling and shipping books and sundry" in "an excellent repurposable computing platform" (Yegge, 2011).

Once the APs found a way to solve the considered problem, they need to retrieve the relevant data. The only point that was raised when searching for data was the fact that the existence of an easy communication in the company is beneficial for the AP. Since there are multiple sources of data, if a low-level developer finds it, then he might have to report it to someone higher than him in the hierarchical system of the company he works for. This report might simply not be listened. The increase in the quantity of available data makes this part more complicated and costly, as explained by the fourth interviewee, but the cost of the data in itself might be cheaper. Once the data is collected, we need to consider a new set of steps increasing the value of the data, and render it usable for the developers:

- Certification: an AP first needs to evaluate how critical the data is, i.e. how much the eventual existence of mistakes is problematic. Moreover, the AP needs to determine if the lack of responsible entity matters, and if the eventual brand recognition is sufficient (I mean, whereas Mozilla Public License explicitly states that their software is provided "without warranty of any kind" (Mozilla, 2014), some might consider that Mozilla's software is reliable). This step differs from the others considering that it is a new one and specific to OD.
- Modification (corrections, cleaning, updates, changes of formats, etc.): before working with the data, the AP needs to make sure that the data are usable. Considering the low quality of the data present on the OD platforms, as discussed in the interviews, raising it, requires money and time.
- Wrapping: the interviewees agreed with the fact that data can hardly be a starting point for the work. However, even once the AP got its idea and knows how to reach its goal, building an application around the data (or plugging the data in an application) might not be a straightforward step. Consequently, wrapping it in a layer of software, shaping it can significantly help the work of the developers.

Once the data is prepared, we can consider that the development process starts. One of the interviewee praised the use of agile methods during this process. Regarding agile methods, what he emphasised was the presence of short iterations and the importance of flexibility (opposed to a carefully planned process), "Responding to change over following a plan" (Beck, et al., 2001). We need to consider that such methods imply a closer relationship between the previously discussed steps and the development. This means that the work on the data is not done when the development starts, and that the data do not need to be perfectly usable to start developing. An iterative process where the AP goes back and forth from the data to the application is the most suitable. It is interesting to relate this iterative process to Donald Schön's "seeing-drawing-seeing" process (Schön, 1992). He explained that "the designer sees what is 'there' in some representation of a site, draws in relation to it, and sees what he/she has drawn, thereby informing further designing" (Schön, 1992). Thus, OD can also be understood as a material with which the designer has a conversation, using the data as an object of design.

The agile manifesto states that communication and individual interactions within the team are essential for a more efficient project (Beck, et al., 2001). This communication can be improved in different ways. Multidisciplinary teams contribute to this goal by limiting the division within the team. In an ideal state, there is not a team composed of a group of designers, a group of

data specialists or a group of developers, but a single team where the pool of talent is dispatched to the different members of the team (by multidisciplinary, I do not mean that specialisation should be avoided). It is also possible to increase communication to a higher scale, in bigger teams or between teams through internal data opening, as discussed in the interview. A coordinated internal opening, with repositories of APIs and an infrastructure built around it, like at Amazon, can increase the internal communication processes (Yegge, 2011).

It is also worth mentioning that the retrieving and treatment of the data parts are not that intertwined with the rest of the development process. Consequently, it is possible to outsource these parts, as discussed in the interview. Actually, some companies have already taken this business opportunity. This point will be elaborated in the next section.

My analysis helped me to identify a key point in the design process: the involvement of people with specific skills. Indeed, in this discussion, I mentioned several times the role of the developer, implicitly spoke about managers when discussing decision-making processes, but I did not mention other significant people whose contribution to the projects should not be disregarded. Firstly, we need to consider the role of the designers. These people are characterised by a creative facet, and while contributing to several different tasks, their first intervention occurs during when the AP is trying to shape an idea. The second group of people that intervene in the design process can be compared to journalists. They participate to the retrieval of the data, find new sources and determine how valuable a dataset is. The third category of people involved in the process is the category of data specialists. They know how to process datasets, how to format them and clean large quantity of data without difficulties. Their presence is essential during the treatment of data, considering that without them, the process would be extremely slower. Lastly, APs need people who are able to describe complex datasets, formalise relationships between entities and create models of data in databases. Whereas a data specialist can have some of these skills, a unique name for this job is XML expert.

This design process can be summarised in the following figure (Figure 3 Design process), which presents the chain of tasks that are needed to be done (and the ones that can be outsourced), the different skills and at what kind of task they would be beneficial.

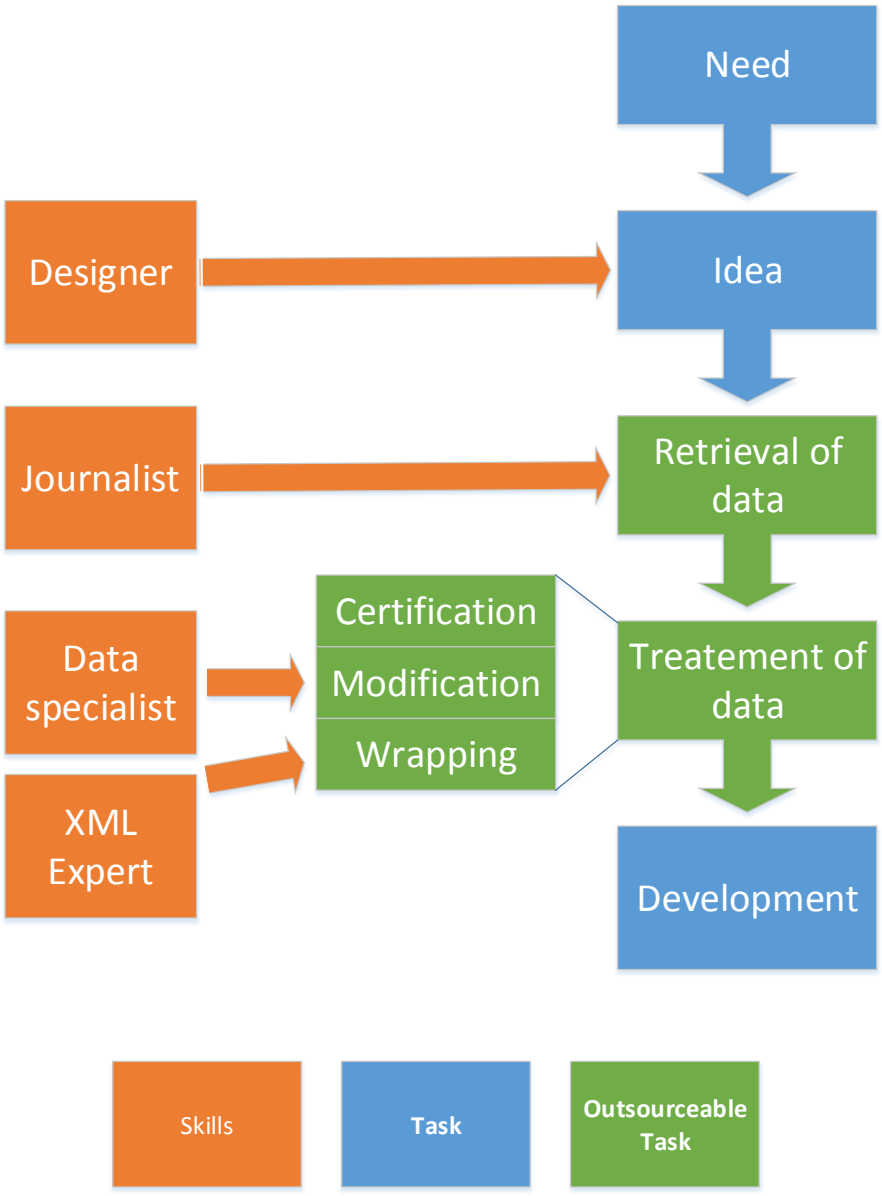

Figure 3 Design process

The ecosystem around the Application Providers

In my analysis and in the previous discussion I started to show how the ecosystem surrounding the APs is changed. The initial and obvious interaction described and discussed since the introduction is the one between the APs and the OD providers. The platforms publish datasets, maintain them and eventually update them. On the other hand, there are the APs that do not have anything to do with the providers, which, from the perspective of the APs, are reduced to repositories of datasets. Initially, the APs retrieve the data from the platforms and then they need to treat them, as explained in the previous section. This perfectly illustrates the duopoly described by the third interviewee, with a group of providers and a group of consumers.

However, I explained how the task that the AP has to do is complex and costly. Interviewees pointed that the platforms provides data, but no service. Consequently, an intermediate service has a certain potential. Such service can act at two different levels (which are not necessarily separated):

- It can take as input data selected by the AP, and then output it in a form that would satisfy the AP's need;
- It can also act as a complete intermediary between the AP and the data platforms, providing the AP with catalogues of data, which can then be customised to suit its need.

Two kinds of entities can provide an AP with this service (I am not considering here the AP who does it on the fly). The first one, regarding the big companies, can be a part of the company. The third interviewee explained how this kind of section of the bank that he works for takes data from many sources as input, and then outputs them to the other sections. Whereas he was not working with OD, it was completely possible to imagine a same method in a structure that relies on multiple OD sources. The second kind of entity takes the form of a company, providing a paid service. Then this business opportunity brings a third category of actors, between the APs and the platforms, companies like OpenDataSoft or Data Publica

Limitations of the data collected

In the two previous sections, I discussed my findings, in order to be able to answer my research question in the next chapter, the conclusion. However, so far, I did not discuss about the flaws of my data, what does it miss and what does it imply on my results. I identified four main limitations of my data. The first one concerns the lack of strong opposition between the presence and the absence of OD in projects, and the three other ones are related with the data gathered from the interviews.

One weakness of my data is that whereas it can explain how to work with OD, I did not always opposed OD to the absence of it. I mean, whereas some points can be essential when working with OD, I am not able to explain how they are more essential when working with OD than in more conventional situations. For instance, I previously presented the interest of agile methods for OD, but are not these benefits the same for every software development projects? It would therefore be interesting to quantify the benefits of these methods in both contexts.

Secondly, I interviewed four participants. In *InterViews: An Introduction to Qualitative Research*, Kvale explains that there is no absolute number of interviews to do, but what important for the

interviewer is to figure when he is not learning anything new from the interviews anymore (Kvale, 1996). In my situation, I quickly started to observe patterns and repetitions through the interviews, and the quantity of new content started to be low during the fourth interview. However, each interview generated some new knowledge. This implies that more interviews would have been beneficial to my data collection. This has not been possible, considering the time limit I had to follow.

The collected data could also have been improved if I was able to come back to the first interviewees after the completion of all the other interviews. Indeed, consequently to my iterative process, the topics of discussion have been improved, their number increased. Moreover, some points, which seemed to be minor ones during the first interviews, were actually more important that what was expected. Consequently, being able to interview the first interviewee again, after the fourth one could have also improved the overall quality of the data.

Lastly, I would have got better quality data by doing my in-depth analysis of the content after and between each interview. Indeed, the idea was to go through all the interviews in a short amount of time, in order to be able to spend more time on an overall analysis. However, this limited the understanding of the topic I built after each interview. The quality of the discussion during the consequent interviews could have been improved by spending more time analysing data between interviews.

Conclusion

The goal of this thesis was to study the effect of OD on APs. In the introduction, I explained how this field of study is relatively new, that it has not been studied much yet. This implies that whereas this thesis does not aim to deliver a complete and in-depth study of the topic, I can contribute to the relevant field in two ways. The first way is by providing the answers to the two research questions of this thesis. I proposed a first approach to this field through my research questions. By answering them, which is the first goal of my thesis, I will improve the existing knowledge on the relationship between OD and APs. The second contribution of my thesis to this topic is done by proposing new questions. Indeed, I explained how this thesis is exploratory. Consequently, its secondary role is also to provide the reader with avenues for possible future research.

Thus, I analysed the impact of OD on APs through two main angles. This divided my research problem into two questions. Consequently, I studied the way the APs work with data and the way they interact with the ecosystem around them.

Through analysis and discussion, this project aims to study the effect of Open Data on Application Providers. Do the Application Providers' software development methods need to be changed In order to work with such data? How does the existence of Open Data shape the ecosystem around the Application Providers?

To answer the first question, I have outlined a model for the design process followed by APs when working with OD. This model identifies a set of tasks that APs need to go through while designing an application based on OD. Moreover, this model also identifies skills that are beneficial to possess when working with OD, stating at which step of the design process they are needed. We can divide these tasks into two categories, the ones that are new, and the ones that are new in the way the APs address them when working with OD. It is common to depict a normal design process as a loop between four steps, requirements planning, design, development and tests. I do not propose any changes to this overall process here. Therefore, these four steps are still valid; however, it is interesting to observe how it is possible to subdivide the development part in a set of several tasks. When looking at the description of the

development process, there is no focus on the existence of data used in this process. It is simply considered as something you use, or generate, while developing. However, when working with OD the developers are confronted with datasets that arrive as is. They do not have any control on these datasets; moreover, they cannot decide what format they would like to choose, what level of precision should be provided, etc. This implies that they have to adapt their work to the data, or the data to their work (depending of the context). This adaptation process is what I presented during this thesis.

Using external data, which is not necessarily warranted by anybody, might be problematic to some organisations. A process of certification or validation is a necessary step when working with external data. Once the OD retrieved, it is often mandatory to clean it and correct it. Indeed, all the interviewees pointed to the fact that most of the datasets downloaded from public platforms need to be cleaned, their formats might need to be changed. Lastly, whereas sometimes the use of data is obvious, its value is not always evident. Wrapping it in a piece of software, shaping it (without changing its content) might render the data more usable. I consider this wrapping task as one of the central ones of my model. It might be the most complicated one to apprehend, however, mastering this task could be the key point for an AP to work with OD in the most efficient way.

I also identified a set of skills, associated to jobs, whose involvement might benefit the development process. The identified jobs, as defined by the interviewees, are the journalist, the data specialist, the XML expert and the designer. These people contribute at different stages of the process. For instance, the journalist is able to retrieve and understand the value of some datasets, while the XML expert can describe and formalise complex relation between entities. Whereas the first one's actions mostly happen during the retrieval of the data, the second one would work more on the wrappers mentioned earlier.

Moreover, I described how and why internal communication is a key point for an AP. There are many different ways to increase the quality of the communication in a company, but the internal opening of its data can contribute to it. Moreover, the opening of data, both internal and external ones, is a way to foster innovation. By opening its data, APs let people with different interests to access it. Then, some of this people might find a way to fulfil one of their needs by using this data. It is possible to relate this process to the democratisation of innovation. Eric von Hippel explained in his book *Democratising Innovation* that lead users are a key point to innovation. Indeed, he argues that people are more likely to find a way to solve a problem they encountered than someone who is not concerned by this problem. Consequently, people can

find solutions to their problems by using the APs' data, whereas the APs might not be able to think about these solutions.

Earlier in this conclusion, I presented a set of tasks that are specific to OD. One of their characteristic is that they do not have to be performed by the AP, or by the team developing the application. It is perfectly possible to outsource these tasks to another department of the company, or to another company. This business opportunity already permitted the emergence of companies providing such services, like OpenDataSoft or Data Publica. Companies like those contribute to the creation of an ecosystem and change the interaction model. Instead of having a simple duopoly with data providers and data consumers, the latter ones, namely the APs, can interact with intermediaries, which retrieve and treat the data. I explained in the introduction that reports consider that OD provides organisations with business opportunities (MESPIR, 2006), and in this thesis, I identified a category of business which are these opportunities.

Now that I summarised my findings, I can propose some clues for further researches:

- During my work, I emphasised the importance of the data wrappers. One of the interviewees even stated that the value comes from the wrappers, and not from the data. Further and more detailed studies of wrapping would be beneficial in order to provide processes and tools for this important and often time-consuming task.
- My thesis does not provide the reader with any technical details. Whereas it is not its goal, since I speak a lot about developers, the proposal of the technical solutions to the problems I evoked would be a plus. For instance, it is possible to dig into the notion of APIs between the data and the developers.
- It would also be beneficial to study the companies I presented, which offer this service built on top of OD. I briefly described those steps and explained that some of them could be outsourced; however, the services offered by those companies are a real life example of what can be done on top of OD, in order to render it as usable as possible.
- I did not provide with many details the action of retrieving data. Creating models for an efficient retrieval process, studying the methods of the "journalists" could be an interesting contribution to the field.
- While being less academic, to design the tools used at the different stages of the design process would be interesting. I showed that there is a plurality of the profiles, that there are many different tasks to accomplish. Consequently, these people are in need for tools to help them in their work.

Therefore, even if there are ways to improve this research or convey new ones, I can state that this thesis is definitely contributing to this field of studies due to the fact that it analyses a relatively new topic. Furthermore, it can be considered as an interdisciplinary research – the study covers points analysed not just in social sciences but also those that are relevant to software development. Additionally, many companies considering working with OD might be interested in getting familiarised better with this topic as well. Thus, I believe that this research can be regarded as innovative, interdisciplinary and even satisfactory for those who seek knowledge and discussion about the issues related to Open Data.

References

Alexopoulos, C., Spiliotopoulou, L., & Charalabidis, Y. (2013). Open Data Movement in Greece: A Case Study on Open Government Data Sources. *17th Panhellenic Conference on Informatics* (pp. 279-286). New York, NY, USA: ACM. doi:10.1145/2491845.2491876

Azevedo, L. F., Canário-Almeida, F., Almeida Fonseca, J., Costa-Pereira, A., Winck, J. C., & Hespanhol, V. (2011, September-October). How to write a scientific paper – Writing the methods section. *Revista Portuguesa de Pneumologia (English Edition), 17*(5), pp. 232-238.

Barriball, L. K., & While, A. (1994). Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing, 19*(2), pp. 328-335. doi:10.1111/j.1365-2648.1994.tb01088.x

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., . . . Thomas, D. (2001). *Manifesto for Agile Software Development*. Retrieved March 3, 2014, from Agile Manifesto: http://agilemanifesto.org/

Berelson, B. (1954). Content analysis. In G. Lindzey, *Handbook of social psychology* (Vol. 1, pp. 488–522). Reading, MA: Addison-Wesley.

Berners-Lee, T. (2009). The Next Web. *TED Global.* Montery, California.

Bernes-Lee, T. (2006, July 27). *Linked Data.* Retrieved February 27, 2014, from World Wide Web Consorsium: http://www.w3.org/DesignIssues/LinkedData.html

Borzacchiello, M. T., & Craglia, M. (2012). The impact on innovation of open access to spatial environmental information: a research strategy. *International Journal of Technology Management, 60*(1), pp. 114-129.

Charalampos, A., Loukis, E., Charalabidis, Y., & Zuiderwijk, A. (2013). An Evaluation Framework for Traditional and Advanced Open Public Data e-Infrastructures. *13th European Conference on eGovernment – ECEG 2013* (pp. 102-112). Como, Italy: Academic Conferences and Publishing International Limited.

Corbin, J., & Strauss, A. (1990, March 1). Grounded Theory Research: Procedures, Canons, and Evaluative Criteria. *Qualitative Sociology, 13*(1), pp. 3-21. doi:10.1007/BF00988593

Davis, T. (2010). *Open data, democracy and public sector reform.* Retrieved February 25, 2014, from http://www.opendataimpacts.net/report/wp-content/uploads/2010/08/How-is-open-government-data-being-used-in-practice.pdf

DB-Engines. (2014, February). *Complete Ranking.* Retrieved February 26, 2014, from DB-Engines: http://db-engines.com/en/ranking

Denscombe, M. (2007). *The Good Research Guide: For Small-scale Social Research* (3rd ed.). Buckingham: Open University Press.

Eberius, J., Braunschweig, K., Thiele, M., & Lehner, W. (2012). Identifying and weighting integration hypotheses on open data platforms. *First International Workshop on Open Data (WOD '12)* (pp. 22-29). New York, NY, USA: ACM. doi:10.1145/2422604.2422608

ENGAGE. (2011, October 19). *Deliverable D7.7.1 – Analysis Report of Public Sector Data and Knowledge Sources.* Retrieved February 18, 2014, from Engage: http://www.engage-project.eu/wp/wp-content/plugins/download-monitor/download.php?id=4

ENGAGE. (2014). *Project Overview.* Retrieved February 18, 2014, from Engage: http://www.engage-project.eu/wp/?page_id=166

Englander, M. (2012). The Interview: Data Collection in Descriptive Phenomenological Human Scientific Research. *Journal of Phenomenological Psychology, 43*(1), pp. 13-35. doi:10.1163/156916212X632943

Franzosi, R. (2004). Content analysis. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao, *Encyclopedia of social science research methods* (pp. 187-190). Thousand Oaks, CA: SAGE Publications. doi:http://dx.doi.org/10.4135/9781412950589.n166

Gillman, B. (2000). *The research interview.* London: Continuum.

Gomm, R. (2004). *Social Research Methodology. A critical introduction.* Hampshire, England: Palgrave Macmillan.

Gurstein, M. B. (2011, February 7). Open data: Empowering the empowered or effective data use for everyone? *First Monday, 16(2).* doi:http://dx.doi.org/10.5210/fm.v16i2.3316

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design, 37*(4), pp. 682-703. doi:10.1068/b35097

Holsti, O. R. (1969). Content analysis for the social sciences and humanities. Reading, MA: Addison-Wesley.

Housden, M., & Crouch, S. (2003). *Marketing Research for Managers, 3rd edition.* Oxford: Butterworth-Heinemann.

Hutchinson, S., & Skodol-Wilson , H. (1992). Validity threats in scheduled semistructured research interviews. *Nursing Research, 41*(2), pp. 117-119.

Kaplan, A. (1943, October). Content Analysis and the Theory of Signs. *Philosophy of Science, 10*(4), pp. 230-247.

Kvale, S. (1996). InterViews: An introduction to qualitative research interviewing. London: Sage.

Latif, A., Saeed, A. U., Hoefler, P., Stocker, A., & Wagner, C. (2009). The Linked Data Value Chain: A Light Weight Model for Business Engeneers. *I-SEMANTICS 2009 International Conference on Semantic Systems*, (pp. 568-575). Graz, Austria.

Le Monde. (2011, May 24). *Tweetdeck racheté par Twitter pour 40 millions de dollars.* Retrieved from Le Monde: http://www.lemonde.fr/technologies/article/2011/05/24/tweetdeck-rachete-par-twitter-pour-40-millions-de-dollars_1526786_651865.html

Lockyer, S. (2004). Coding Qualitative Data. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao, *Encyclopedia of Social Science Research Methods* (pp. 138-139). Thousand Oaks, CA: SAGE Publications. doi:http://dx.doi.org/10.4135/9781412950589.n130

Mahoney, C. (2004). Probing. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao, *Encyclopedia of Social Science Research Methods* (pp. 872-873). Thousand Oaks, CA: SAGE Publications, Inc. doi:http://dx.doi.org.zorac.aub.aau.dk/10.4135/9781412950589.n760

Mason, J. (2004). Semistructured interview. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao, *Encyclopedia of social science research methods* (pp. 1021-1022). Thousand Oaks, CA: SAGE Publications, Inc. doi:http://dx.doi.org/10.4135/9781412950589.n909

MESPIR. (2006, June). *MESPIR Measuring European Public Sector Information Resources.* Retrieved February 18, 2014, from ePSIplatform: http://www.epsiplatform.eu/ezpublish_media/MEPSIR%20Final%20Report.pdf

Mozilla. (2014). *Mozilla Public License Version 2.0*. Retrieved from Mozilla Project: http://www.mozilla.org/MPL/2.0/

Nay-Brock, R. M. (1984). A companson of the questionnaire and interviewing techniques in the collection of sociological data. *Australian Journal of Advanced Nursing, 2*(1), pp. 14-23.

Newton, N. (2010). *The use of semi-structured interviews in qualitative research: strengths and weaknesses*. Retrieved March 13, 2014, from Academia.edu: https://www.academia.edu/1561689/The_use_of_semi-structured_interviews_in_qualitative_research_strengths_and_weaknesses

Obama, B. (2009, January 20). *Transparency and Open Government*. Retrieved from The White House: http://www.whitehouse.gov/the-press-office/transparency-and-open-government

ODI. (2014). *What is open data?* Retrieved February 27, 2014, from Open Data Institute: http://theodi.org/guides/what-open-data

OECD. (2004, January 29-30). Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communique. Retrieved from OECD: http://www.oecd.org/science/sci-tech/sciencetechnologyandinnovationforthe21stcenturymeetingoftheoecdcommitteeforscientificandtechnologicalpolicyatministeriallevel29-30january2004-finalcommunique.htm

Olin & Uris Libraries. (2012, September 18). *Distinguishing Scholarly Journals from Other Periodicals*. Retrieved March 11, 2014, from Olin & Uris Libraries: http://olinuris.library.cornell.edu/ref/research/skill20.html

O'Neil Green, D. (2008). Categories. In L. M. Given, *The SAGE Encyclopedia of Qualitative Research Methods* (pp. 72-73). Thousand Oaks, CA: SAGE Publications. doi:http://dx.doi.org/10.4135/9781412963909.n40

Open Definition. (2005, August). *Open Definition*. Retrieved from Open Definition: http://opendefinition.org/

Open Knowledge Foundation. (2014). *About*. Retrieved from Open Data Index: https://index.okfn.org/about/

Open Knowledge Foundation. (2014, February 13). *Countries.* Retrieved February 13, 2014, from Open Data Index: https://index.okfn.org/country/

Open Knowledge Foundation. (2014). *Our Vision*. Retrieved from Open Knowledge Foundation: http://okfn.org/about/our-vision/

Schön, D. A. (1992, September 01). Designing as reflective conversation with the materials of a design situation. *Research in Engineering Design, 3*(3), pp. 131-147. doi:10.1007/BF01580516

Smith, L. (1992). Ethical issues in interviewing. *Journal of Advanced Nursing, 17*(1), pp. 98-103.

Stebbins , R. A. (2008). Exploratory Research. In L. M. Given, *The SAGE Encyclopedia of Qualitative Research Methods* (pp. 328-330). Thousand Oaks: SAGE Publications, Inc.

Stebbins, R. A. (2006, October). Concatenated Exploration: Aiding Theoretic Memory by Planning Well for the Future. *Journal of Contemporary Ethnography, 35*(5), pp. 483-494. doi:10.1177/0891241606286989

Tammisto, Y., & Lindman, J. (2012). Definition of Open Data Services in Software Business. *International Conference on Software Business* (pp. 297-304). Cambridge, MA, USA: Springer.

The Economist. (2013, March 18). Open data: A new goldmine. *The Economist*, 73.

Treece, E. W., & Treece, J. W. (1986). *Elements of Research in Nursing.* St. Louis: Mosby.

University of Dayton. (n.d.). *Exploratory Research.* Retrieved March 10, 2014, from University of Dayton: http://campus.udayton.edu/~jrs/tools/notes/exploratory%20research.pdf

von Hippel, E. (2006). *Democratizing Innovation.* Cambridge, Massachusetts: The MIT Press.

Waxer, C. (2013, March 24). *Government open data proves a treasure trove for savvy businesses.* Retrieved April 1, 2014, from Computerworld: http://www.computerworld.com/s/article/9246970/Government_open_data_proves_a_treasure_trove_for_savvy_businesses

Yegge, S. (2011, October 11). *Stevey's Google Platforms Rant.* Retrieved from Google Plus: https://plus.google.com/+RipRowan/posts/eVeouesvaVX

Zuiderwijk, A., Jeffery, K., & Janssen, M. (2012). The Potential of Metadata for Linked Open Data and its Value for Users and Publishers. *eJournal of eDemocracy and Open Government, 4(2).*

Appendix

A - Interview Guide n°1

Interviewee: Software developer who worked with Open Data (OD) on a six months project, and participated to a hackathon.

Theme: Open Data platform

- What do you think of looking for information on these platforms?
- Did you encounter any problems?
    - Finding the information
    - Way information is proposed (format of the file, quality of the data, etc.)
- What do you think of cleaning the datasets, and giving them back to the public?

Theme: Linking datasets

- What benefits do you see from relating datasets?
- How do you usually look at datasets having points in common?

Theme: Internal and external data

- What differences do you see between working with "real OD" and data produced by others services of your company?
- How useful is the possibility to have some sort of support from people working with the data?

Theme: Standards and formats

- How do you dealt with different standards, from your organisations ones, or the ones from external sources (OD)?
- Do you see a shift or an interest in a shift from internal standards to more common (open?) ones?

Theme: Technical constraints of software development

- What change of methodology do you think are important when working with OD?
- Do you see benefits from using agile methodology?

- What form of agile methods should be used?
- Does using OD imply changes regarding the reliability of the sources?

Theme: Open Data as a trigger to innovation

- How do you think OD enables people to innovate?
- How can OD give new ideas?
- How can OD makes ideas possible to do?
- What benefits could you see for a company to free some of its data?
    - Federate a community
    - Outsource some innovation
- What risks could you see for a company to free some of its data?

It could be interesting to oppose the ideas of OD helping people to get more idea, to OD just letting people realise their ideas.

Theme: Changes to the organisation

- Do you think that OD requires the structure of the organisation you are part of to change (and how)?
- How being able to work with different kinds of data (incl. ones not "common" for the core of the organisation) can shape this "core business"

B - Interview Guide n°2

Interviewee: Software engineer who participated to a hackathon and works with Twitter's data.

Theme: Open Data platform

- What do you think of looking for information on these platforms?
- Did you encounter any problems?
    - Finding the information
    - Way information is proposed (format of the file, quality of the data, etc.)
- What do you think of cleaning the datasets, and giving them back to the public?

Theme: Linking datasets

- What benefits do you see from relating datasets?
- How do you usually look at datasets having points in common?

Theme: Internal and external data

- What differences do you see between working with "real OD" and data produced by others services of your company?
- How useful is the possibility to have some sort of support from people working with the data?

Theme: Understanding the data

- How would you describe the process of understanding the meaning of data you did not create?
- How do you try to determine what is doable with data?

Theme: Standards and formats

- How do you dealt with different standards, from your organisations ones, or the ones from external sources (OD)?
- Do you see a shift or an interest in a shift from internal standards to more common (open?) ones?

Theme: Putting an API between the data and the developers

- What do you think of a REST-like API after the data sources, to abstract the files?

- How do you imagine such APIs? Simply Restful, or higher level?

Theme: Technical constraints of software development

- What change of methodology do you think are important when working with OD?
- Do you see benefits from using agile methodology?
- What form of agile methods should be used?
- Does using OD imply changes regarding the reliability of the sources?

Theme: Open Data as a trigger to innovation

- How do you think OD enables people to innovate?
- How can OD give new ideas?
- How can OD makes ideas possible to do?
- What benefits could you see for a company to free some of its data?
    - Federate a community
    - Outsource some innovation
- What risks could you see for a company to free some of its data?

It could be interesting to oppose the ideas of OD helping people to get more idea, to OD just letting people realise their ideas.

Theme: Changes to the organisation

- Do you think that OD requires the structure of the organisation you are part of to change (and how)?
- How being able to work with different kinds of data (incl. ones not "common" for the core of the organisation) can shape this "core business"?
- What about the company's culture? Giving more latitude to the employee through a bigger access to the data might empower them.

C - Interview Guide n°3

Interviewee: Software engineer who participated to a hackathon.

Theme: Open Data platform

- What do you think of looking for information on these platforms?
- Did you encounter any problems?
    - Finding the information
    - Way information is proposed (format of the file, quality of the data, etc.)
- What do you think of cleaning the datasets, and giving them back to the public? ➔ speaking of the "short-term", compared to OSS

Theme: Linking datasets

- What benefits do you see from relating datasets?
- How do you usually look at datasets having points in common?

Theme: Internal and external data

- What differences do you see between working with "real OD" and data produced by others services of your company?
- How useful is the possibility to have some sort of support from people working with the data?

Theme: Development process

- If you should describe the development process from the beginning to the end, how would you do it?
- I can suggest: Idea for application ➔ retrieve data ➔ understand data ➔ make data usable + cleaning data ➔ creation of the application
- Would you suggest something else?

Theme: Standards and formats

- How do you dealt with different standards, from your organisations ones, or the ones from external sources (OD)?
- Do you see a shift or an interest in a shift from internal standards to more common (open?) ones?

Theme: Understanding the data

- How would you describe the process of understanding the meaning of data you did not create?
- How do you try to determine what is doable with data?
- An interviewee made a parallel with the role of journalist, having access to data and try to figure if something is doable from it, what is your position on this question?

Theme: Putting an API between the data and the developers

- What do you think of a REST-like API after the data sources, to abstract the files?
- How do you imagine such APIs? Simply RESTful, or higher level?

Theme: Technical constraints of software development

- What change of methodology do you think are important when working with OD?
- Do you see benefits from using agile methodology?
- What form of agile methods should be used?
- Does using OD imply changes regarding the reliability of the sources?

Theme: Open Data as a trigger to innovation

- How do you think OD enables people to innovate?
- How can OD give new ideas?
- How can OD makes ideas possible to do?
- What benefits could you see for a company to free some of its data?
    - Federate a community
    - Outsource some innovation
- What risks could you see for a company to free some of its data?

It could be interesting to oppose the ideas of OD helping people to get more idea, to OD just letting people realise their ideas.

Theme: Changes to the organisation

- Do you think that OD requires the structure of the organisation you are part of to change (and how)?
- How being able to work with different kinds of data (incl. ones not "common" for the core of the organisation) can shape this "core business"?

- What about the company's culture? Giving more latitude to the employee through a bigger access to the data might empower them.

D - Interview Guide n°4

Interviewee: Software engineer who participated to a hackathon.

Theme: Open Data platform

- What do you think of looking for information on these platforms?
- Did you encounter any problems?
    - Finding the information
    - Way information is proposed (format of the file, quality of the data, etc.)
- What do you think of cleaning the datasets, and giving them back to the public? ➔ speaking of the "short-term", compared to OSS

Theme: Linking datasets

- What benefits do you see from relating datasets?
- How do you usually look at datasets having points in common?

Theme: Internal and external data

- What differences do you see between working with "real OD" and data produced by others services of your company?
- How useful is the possibility to have some sort of support from people working with the data?

Theme: Development process

- If you should describe the development process from the beginning to the end, how would you do it?
- I can suggest: Idea for application ➔ retrieve data ➔ understand data ➔ make data usable + cleaning data ➔ creation of the application
- Would you suggest something else?

Theme: Standards and formats

- How do you dealt with different standards, from your organisations ones, or the ones from external sources (OD)?
- Do you see a shift or an interest in a shift from internal standards to more common (open?) ones?

Theme: Understanding the data

- How would you describe the process of understanding the meaning of data you did not create?
- How do you try to determine what is doable with data?
- An interviewee made a parallel with the role of journalist, having access to data and try to figure if something is doable from it, what is your position on this question?
- Designer?

Theme: Putting an API between the data and the developers

- What do you think of a REST-like API after the data sources, to abstract the files?
- How do you imagine such APIs? Simply RESTful, or higher level?

Theme: Technical constraints of software development

- What change of methodology do you think are important when working with OD?
- Do you see benefits from using agile methodology?
- What form of agile methods should be used?
- Does using OD imply changes regarding the reliability of the sources?

Theme: Open Data as a trigger to innovation

- How do you think OD enables people to innovate?
- How can OD give new ideas?
- How can OD makes ideas possible to do?
- What benefits could you see for a company to free some of its data?
    - Federate a community
    - Outsource some innovation
- What risks could you see for a company to free some of its data?

It could be interesting to oppose the ideas of OD helping people to get more idea, to OD just letting people realise their ideas.

Theme: Changes to the organisation

- Do you think that OD requires the structure of the organisation you are part of to change (and how)?

- How being able to work with different kinds of data (incl. ones not "common" for the core of the organisation) can shape this "core business"?
- What about the company's culture? Giving more latitude to the employee through a bigger access to the data might empower them.

E - Summaries of the three first interviews

First interview

The first subject did not work with OD coming from platforms (he mostly worked with OpenStreetMap), but he looked at these platforms and has been able to make his opinion on the matter. He regretted that most of the times, the platforms provide the data as is, in raw files. The developers have to do everything, from seeing how the file is constructed to render it usable during the development of the application (by transferring the data in a database, or creating an extra-layer in the application to read the files). He said that a RESTful API giving access to the resources hosted in the files would be a way better solution. Such solution would facilitate the work of the developer. Moreover, during the second theme (linking datasets), he raised the point that the linked data method is often (or can easily be) implemented in RESTful APIs.

During the discussion about internal data, he stated that having direct access to the producer of the data – who is someone from the company – helps to the understanding of the data. This notion of the need to understand data, before using it misses from my list of themes, and should be explored in the following interviews.

The third and last big point raised by this interview is the fact that he considered that "working with OD does not change the structure of the company, but to work efficiently with OD, some structures are better than others". He especially mentioned structures with an efficient top-down communication process, i.e. "flat hierarchies". He considered that without such structure, the company, as a whole, would not be able to be aware of the existence and uses of available and free data. We can illustrate that with a developer finding an interesting dataset, which could permit the creation of a new product or feature. However, the company would never use it, since the top management would never hear of it.

He also stated that flattening the structure of the company would benefit its culture, making people feel empowered by using OD (both internal and external). To explain his point, he made a comparison with the effect of OD in countries, giving the opportunity to citizen to contribute to the society.

Second interview

The second interviewee participated in two hackathons using OD. Consequently, he used OD platforms. He said that there is plenty of data available on these platforms, the quality seems to be good, but it is hard to find something usable. Moreover, the formats of the files did not

correspond to his expectations, and the data had to be cleaned. It is hard to see the interest for people cleaning and republishing data.

Linked data and URIs is an interesting idea, but he does not believe it is doable.

Working with internal data or external data has similarities, and opening data internally can be a first step toward an external opening.

We then discussed about the new topic, which is the understanding of data. He raised the point that it is possible to compare such work with the one of a journalist. We have seen recently many journalistic initiatives focused on the data, its visualisation, and explanation.

Formats are indeed an issue, however, it is never possible to work with raw data, it has to be parsed and put in databases. It is unlikely to reach a consensus with universal formats, since there is more and more data, but we might start to use third part providers which will converting the different format to one (for instance, TripAdvisor gathers the data about restaurants, and then propose it, through an API for instance).

Putting an API on top of the data has benefits, by creating high abstract models of the data, possibly with linked data inside, however, this add a layer of development. Therefore, it do not suit all projects. Moreover, the interviewee usually prefers to work directly with a database or the files, in order to have a full control of the data, but he understands that web developers might be more comfortable with a REST API.

Working with OD makes some development easier, mainly by saving time and money, but it also imply that the data has to be trusted, and therefore, verified.

OD has a positive effect on innovation, the data permitting the imagination of new usages. Opening data externally can have positive consequences on the image of the company, and permit the outsourcing of innovation (which can then be copied/bought). However, this can force more transparency, and facilitate the work of competitors.

Lastly, we discussed on the relation between OD and the organisation of companies. Firstly, OD implies that companies need to have in their pool of skill a data specialist, i.e. someone able to handle, aggregate and parse the data. Secondly, having a good communication within the company to be able to catch the opportunities to work with new data is important. Lastly, opening data internally might have a positive effect on the culture of the company.

Third interview

The third interviewee participated to a hackathon, and worked with OD stored on OD platforms. He observed that the quality of the data is low, for two main reasons: first, the authors do not really know how to describe their data, and often use bad formats to do it. Secondly, the datasets stored are not that interesting, having generally specific and limited uses, or being outdated. He considers that a good application is an application that solves daily problems; consequently, old data is not going to be highly valuable.

He then starts to speak about his job in a bank (on the financial side). Their main issue with data (in general, not OD) is to be able to describe it correctly, to make it usable, to makes the models representing complex data relationships. Therefore, jobs such as XML specialist (people describing the structure of data) or ETL specialists (Extract, Transform and Load) are highly valuable. In his company, they get several different streams of data (both internal and external), and need to output data in homogeneous way. This process increases highly the value of the data, making it usable. They do not change the data, but shape it. Financial companies work with data, which is more or less the ideal state that OD could reach. Indeed, everything is described in a proper and unique way; organisations have access to all the real time information about the market and the companies. He then expand his views on the fact that the value is in the way to present the data, than the data in itself, since it is actually often possible to get access to data online (for instance, public transport schedules, through aggressive web scrapping).

For him, a collaborative OD is not likely to happen, like for the web, there will be producers and consumers, it is two different jobs, and two different skills. Moreover, such collaborative process would require an appropriate platform, which would be too complicated to make.

When discussing the concepts of linked data, he told me that he agrees on the interest, and that the idea is really close to what they are doing in the financial markets, using unique identifiers for institutions, stocks, or other trade objects, through the system of ISIN (International Security Identifier Number) and LEI (Legal Entity Identifier). This dramatically increases the creation of value.

When asked about the development process, he considered that in general the idea comes first, and then we look for data (open or not) to satisfy a need. Therefore, getting an idea from data is usually more useful to display a technology, skills, for a portfolio, but not much to create a real-life service.

We finally fell back on the fact that the value is in the wrappers around the data, and consequently, data specialists are the key to work with any kind of data, rather than people able to find the data.

F - First interview

- 0:00, chitchat, explanation of the thesis, and how the interview will work
- 5:25, beginning of the interview, introduction to theme of platforms
- 5:55, never used these platforms, just had a look
- 6:15, why never used? Was not looking for anything specific, looked for "curiosity", did not find interesting stuff, and reuse did not seem "evident"
- 6:50, expanding. There's a lot of data, but it's just raw data, you need to do everything by yourself, it is not a service
- 8:15, quality of the data, presentation? No homogeny.
- 9:50, "hallucinated when seen the level of detail of the data"

- 10:30, presenting the theme of Linking Datasets
- 11:50, what opinion on that? Don't know about linking within "repos"
- 12:30, explaining that linked data exists in restful, and it's great
- 12:50, "a single access point", and then working from this data
- 13:25, "platform agnostic"
- 13:50, speaking of "non-official platforms"
- 15:50 used the possibility of navigating from a data through another data with REST APIs, but not with platforms

- 16:30, internal open data
- 18:30, more or less same idea, working with OD and internal OD
- 20:20, interesting to have access to a "support", firstly (less important), accessing the data
- 20:40, working with complex data: helping to interpret the data, and mastery of the domain of the data, on how to get the meaning of it
- 22:00, usually easier to working with internal data, it is "related to the needs of the company", and benefits from standardisation

- 24:10, working with different standards

- 24:30, balancing between the needs and the standard, ideally, go for the standard
- 25:20, would convert if has time
- 27:10, having to balance with co-existing standards

- 1:24, software development
- 2:50, the work is probably more important on the side of the entity opening data than the one using it. The "company opening data does more work than the one using it"
- 4:30, reliability of the data? OSM is precise. He believes that it has been proved that Wikipedia is more precise than the Encyclopaedia Britannica.
- 7:30, collaborative OD? Could be a good idea, but is not sure there are examples of this. Make a comparison with FOSS, saying that we know it's beneficial, so could be the same thing with OD

- 9:05, OD and innovation
- 9:30, bringing new ideas, or realising ideas?
- 10:10, completely agrees, both are complementary
- 11:35, there are plenty of examples of companies/business built on top (and just with) OD
- 12:55, innovation by providing a better service than a company from its data (Capitaine Train)
- 0:10, appropriation of the data through OD
- 0:25, by releasing data, we can create new business/usages from the data, even when not expected by the producer of the data
- 2:50, benefits from opening data? Good image for the company, transparency etc.
- 2:25, federating a community? Yes, creation of stuff "hooked on the business", these developers might be evangelists
- 4:30, why not opening it? If it is confidential, of course not open. OD does not mean "let's open everything", it imply that we decide of what to open, how to open it, what format, etc.
- 7:00, discussion about legal constraint, but usually (at least in France) there are public administrations ensuring that databases of people are ok for instance

- 9:00, “it’s not OD which will change the structure of the company, but the structure will affect the efficiency of the work with OD”
- 9:25, for instance, in really hierarchical structure, without good communication, hard for someone “discovering some data” to really use it ➔ needs to be able to tell the manager of the existence/potential of OD
- 11:10, the structure of the company will facilitate the internal opening of data
- 11:30, less sure if this same idea of good structure will facilitate the opening of data to external public, possible ➔ comes with the idea of “creating usages from data”
- 12:45, this API/opening of data might be just a trend
- 13:40, an internal API approach in the company is interesting thing, always with the “creating usages”
- 15:00, having access to data not-related to the core business might be really helpful, to hint, to guide the business in the same thing
- 16:00, again, having access to not related data can create new services from the company (“new usages”)

- 19:30, “flattening the company”, just like OD let the citizen contribute, the employee could involve himself in the life of the company

- 20:10, Done. Chit chat

G - Second interview

- 00:00, chit chat
- 1:30, introduction
- 7:10, first theme
- 8:00, used official platforms + CKAN
- 8:20, CKAN, a lot of data, really generalist, so hard to find something interesting
- 8:35, on PACA platforms, a lot of stuff, interesting, but did not really used interesting
- 9:15, evaluation of the quality of the data is good
- 9:30, need for a community
- 10:20, was looking for RDF files (semantic data), did not find what he was looking for
- 11:30, republish cleaned data ➔ cleaning data is important
- 12:00, data-publica is a company cleaning/analysing public data and selling it to companies
- 13:40, big difference source code and data. Updated source code will lives through the updates, if data is updated every year, the "cleaning" won't be reusable next time
- 14:40, interesting to give the cleaned data to the administration, not just to the "public" through platforms

- 16:00, linked-data/meta-data
- 16:40, URI essential to semantic web
- 17:05, URI in OD would be "magical", but impossible ➔ people without any links together
- 18:20, geographical data: URI not necessary, could be nice
- 21:45, did not navigate "through data"

- 24:30, internal data
- 25:40, some companies don't want to open data to the society, but can do it internal ➔ risk that they just open internally, and not externally
- 26:30, agrees with the fact that doing it internally might be a first step to the externalisation of data
- 28:20, external and internal data is more or less similar

- 29:40, internal API is safer than a file, since it's easier to "steal the file" and give it to another company

- 31:15, understanding of data
- 33:20, to understand and analyse the data is the role of a "journalist", starting to get common in some newspapers, Le Monde (*décodeur*), etc., data journalists?
- 37:20, not convinced of the usefulness of a designer during this process, more the role of the journalist
- 38:15, "veilleur" ➔ scouting in a company context, and has a real power of decision, knowing the trends

- 39:45: formats and standards

- 40:30, essential to parse and treat all the dataset, to put them in the database
- 41:30, impossible to work on the raw data, if you want to use more than just one dataset
- 43:00, complicated to reach the "one format", since there is more and more data
- 45:00, hard to reach a consensus on the format of data in public/regional administrations ➔ it will take time, but it should happen
- 46:00, instead of looking directly to the public sources, ask a third party actor for a "curated data", from tripadvisor for instance, or the "national portal" to have a formatted data

- 2:05, API?
- 4:20, add a layer of software/development
- 4:50, useless if a low number of files
- 5:00, might be nice if a lot of files, or if a file is too big
- 5:30, beneficial to be able to plug to someone else's API, up to date, etc.
- 6:30, API useful if normalisation of the data, useless if just an abstraction of the file system
- 10:15, URI/Linked-Data good, but need another layer of work

- 11:00, we can't directly compare CSV and API, since the API requires extra code
- 11;30, depends on the needs
- 11:50, like to have a direct access to a database

- 12:50, Software development
- 13:40, possible to do something quicker and more easily (get all the cities of a country)
- 14:40, effects of people at low levels in the company, creating "manually" the data
- 15:30, managers less likely to trust the data
- 17:15, working with OD imply that you will check the validity of the data, not with internal data
- 18:45, doubts that working with OD will really change the companies

- 20:40, innovation
- 21:20, reduction of costs for a company
- 22:00, agrees with the fact that it create innovation, now that access to data "imagination of new usages"
- 24:00, so many examples of apps working on top of free/open data
- 27:40, open source software access to Wikipedia ➔ DBPedia
- 31:20, opening data?
- 31:45, benefits to image
- 32:15, companies doing public hackathons, with data "opened just for the hackathon" ➔ examples of data that companies which could open some data
- 36:40, creation of new usages, that the company will be able to buy/copy
- 40:00, risk of facilitating the work of competitor, or lowering the "price to enter the market", if it's a data all competitor already have/need
- 42:30, impose a transparency, risk of discovering some lies "we have 1k visitors per month, but we apply for a subvention for 2k visitors per months", publishing stats might be a risk for a company

- 45:45, changes to organisation

- 47:20, need to have a "data engineer", someone able to handle the data, aggregate them, parse them, etc.
- 49:45, necessary that the person who "finds the data" is listened in the company
- 50:20, the will to work with OD does not always come from the bottom, it is something known now
- 51:00, executives might be afraid to work with another company's data
- 54:30, using OD to diversify the business, or create partnerships with other companies ➔ opening some data to this company?
- 55:30, opening internal data in a company, helps to know more about the other department: useful if there is communication, not the only solution ⇔ company social network
- 57:00, people within the company might find usages from internal data
- 58:00, internal OD as ONE way to improve the dynamic of the company
- 59:00, people having the monopoly of some knowledge in a company won't publish it, risks of "losing their job"

- 1:02:00, making internal hackathon in a company improves the communication within the company
- 1:03:00, might be hard for the culture of the company

H - Third interview

- 0:00, introduction
- 7:00, OD platforms
- 7:40, had the opportunity to "search, analyse a datasets, use a repository, and create an idea of service, for a hackathon"
- 8:10, observed that the actors providing with the data don't "know how to describe the data". They want to contribute, but do not have the tools to transfer what they own. Usually, the file format is not the god one (RTF when not useful)
- 9:20, the actors balance between two positions, following the minister's dynamic, and valorisation of their property. For the developer, wanted to get the schedule of the museums. But datasets are outdated: "top 100 of 1993 movies"
- 11:40, guess they don't give some datasets (schedules) because they sell it, so monetary value. "economic interests start when open data stops"
- 12:30, in brief, bad format, outdated, people not really sensible to the problems, and economic interests
- 13:20, to create value, need to solve daily problems. So need to real time data. For instance, transports in Parisian region: several millions of people spending more than one hour in the transports every day, saving them 10 minutes would be great, but this data is not open, it has a price
- 14:15, the idea behind open data: you don't make money of this data, and let people valorise it
- 15:00, they can't access to let it for free, for instance Météo France, whereas in the USA it's free
- 16:00, about correcting data?
- 17:10, speaks about his job: works in a bank, with two important things, market and reference data
- 17:30, big difficulty now, to describe what we use from a business perspective: need people able to describe, make models of data (in databases for instance), for instance, legal structure between companies, correlation between probabilities of consumer defaults, etc.
- 18:20, XML expert: able to make models of data which does not fit in a 2D table
- 18:50, top jobs: "ETL" specialist, extract, transform and load

- 19:15, his company has just one problem, to make these models. They centralise data, and then dispatch it ➔ internal open data, the company does not sell its data from one service to another one
- 19:50, heterogeneous data inputted, homogeneous data outputted ➔ creation of value. It's not a transformation of data, just changing its shape, facilitate reutilisation
- 20:40, the idea of the financial markets is "open data", everybody get the information about the market's state
- 21:40, ideal state, efficient-market hypothesis
- 22:30, we can imagine that with a real OD world, where every economic agent can access the information in real time, we can create value, like in financial markets
- 22:45, they already do that in financial market, since they know there's money to do, we still need to reach this level with OD
- 23:00, it's starting to converge, for instance hedge funds mixing data, doing a regression between a stock exchange performance and meteorological data, observing sport bet rates, etc.
- 23:45, interest of OD is to valorise data, create a wrapper of it, without it, it's impossible to work and to make money
- 24:10, gives a personal example: scrape the time schedules of a school, output them in an ICS format, then easier to build something on it, and people are more on time, etc.
- 24:40, expect an "open data guerrilla", where people scrape the data online, making the data open for their uses when it's not available
- 25:25, for instance, you can fight yield management by "finding their formula". For instance, airplane companies, no transparency on the prices. Some people analyse the data, and valorise it
- 26:30, some companies fight against the easy accessibility to their data, for instance, pagesjaunes.fr do not default displays phone number anymore, needs to click on a button to get it ➔ prevent the people who were pumping the data in an aggressive way
- 27:50, people in ~monopolistic situation fight to protect their data, since their business models is the data, and not really services on top of it
- 28:50, ideally tomorrow we'll describe flux and entities
- 29:30, speaks of people creating services obfuscating the openness of the data
- 32:30, OD will be the death of the monopolistic jobs

- 37:00, about correcting data: it is two different jobs, there are specialists on the data, who produce it, and then people using it.
- 38:20, creating a community version of OD, would require a platform like OD. This would imply that the producer of the data would also be the mediator/validator of the corrections of the data
- 39:00, we'll stay for a while with a duopoly, expert/producer and user
- 39:40, Wikipedia is not a good example, they designed it a collaborative platform, before being a data platform
- 40:40, producing and maintaining the data is already complicated/expensive
- 41:15, like the internet, a small minority of people producing/publishing, the vast majority just consume
- 43:00, data is used, opened or not, people always find a workaround
- 44:30, there will be mainstream OD, with official licences, etc. and people doing OD without knowing it, remixing some data and putting it on internet. Something not protected is *de facto* open
- 47:00, when there is a need, people will find ways to work with closed-data (falling back to this calendar stuff that hundreds of students in the school used)

- 48:00, linked data
- 50:45, symbology, a tool to describe, give a primary key to financial instrument (a stock, an obligation, etc.)
- 51:30, every country has an agency creating NSIN (National Securities Identifying Number) to describe these instruments outputted in the country.
- 52:00, financial institutions need to pay the agencies to have access to the primary keys. Without the keys, impossible to exchange between different institution/databases
- 53:20, unique description releases of any ambiguity
- 54:00, the ISIN (international) is free (not in the USA)
- 54:30, through ISIN/NSIN we can describe what is exchanged, but not who exchanges
- 54:50, every entity has now an LEI (Legal Entity Identifier). To be efficient, need to describe.

- 55:30, Bloomberg created an open symbology (BSYM) ➔ they propose for free their own convention to describe everything (initially, need to use tickers (character of strings), by Reuters or Bloomberg, for instance used in a contract)
- 56:50, with the BSYM, Bloomberg propose a free and open way to map their internal tickers to an open one
- 58:40, another example of primary keys, SWIFT/BIC/IBAN to describe bank accounts from one bank to another one ➔ value creation; same thing for phone numbers, or for the Single Euro Payments Area (SEPA, improving the efficiency of cross-borders payments in Europe)
- 01:00:00, comparison with ICQ/BBM numbers, unique identifiers of people for using a service. However, we look for using only one identifier, not one per service (e.g. Google account/Facebook, email address, phone number, etc.)

- 01:02:00, internal data?
- 01:03:40, if the internal data is not really usable, won't be used
- 01:04:00, not much differences between external/internal data with such an advanced formalisation process

- 01:05:00, process? during the hackathon, was searching for the "less worst" dataset
- 01:07:00, comparison between the two different chains?
- 01:08:30, driven by needs. Not by the data. They need some data, they will look for it. If it is open, it is good (saving time/money), if not workaround.
- 01:09:40, the data ➔ idea process is graphic designers/portfolio/proof of concept stuff, for happy few, not a real creation of value
- 01:10:40, what matters, is people trying to solve a problem, and look for data, open or not
- 01:11:30, for instance Xavier Niel (founder of Free/Illiad, French telecoms) pumped a phone directory to create the first French reverse directory (get a name from a phone number)
- 01:12:00, blocking the data is not really useful, people will get through it, so a clever entity will control the openness

- 01:17:10, internet produces and produces data. How to find the utility/value? First, crowdsourcing, for instance, Reddit is an example of people defining (through votes) what is important and what is not, data needs to be sorted by humans
- 01:18:30, is the "good" data the one used by the people?
- 01:19:30, Bayesian inference to determine the correlation between several data
- 01:20:20, loyalty programs: the companies use them to get data about the consumers, and know how to sell more ➔ it's just about statistics
- 01:23:00, going forward the "data specialists"
- 01:26:00, a better access to data is a competitive advantage
- 01:27:30, same thing in companies, people subject of the software, and those able to master/dominate it
- 01:28:30, culture of the data in companies
- 01:31:00, how about journalists in companies?
- 01:35:30, the goal is to find people able to render the data in a useful format. The difficulty won't be to find data, but make it useful. Shaping is what matters.

- 01:40:00, about structure?
- 01:40:00, need to have the internal skills, the people expert in XML, in data, able to shape data in a useful way, need to produce the good data cubes. Even with evolution of the technology, need to have the specialists
- 01:44:00, chit chat
- 01:46:00, what is essential is to have data specialists

I - Fourth interview

- 8:00, beginning of interview
- 9:00, often the data is not as we would expect it, and then require a "pre-treatment" (see data-publica)
- 11:00, bigger company, working with big data for Business Intelligence, starts to work on the creation of interfaces to structure/format data more easily and use it
- 14:00, collaborative data correction? Why not. But problem: when is the data integrity assured
- 16:00, after integrity, problem of legitimacy: the license needs to authorise changes to data

- 21:30, why not, does it answer to a real need? Discussion about the centralisation of data. What about scalability?
- 23:45, searched for related datasets? No, we're not there yet.
- 24:30, speaks about when releasing data

- 27:00, describes how his company makes data available internally, through "interface contracts", defining everything (structure of the data, way the data is accessed, charges, etc.)
- 28:00, now going through enablers, "api-like"
- 30:30, complicated to create them once the service is created, cheaper/simpler if everything is done in the same time
- 31:30, still really simpler to just "borrow" a file in the company, when no need for any updated data
- 32:30, describe how they started to work with internal data, and made it available, with the beginning of the internet
- 38:30, experience with external/internal data is more or less the same, really rare that a company use the same technology/format/data/etc. from one service to another one. By definition, the needs are different, so the services can't work the same way
- 40:00, however, on a "big data perspective", they try to work on the harmonisation of it, better structure, structured the same way, with one access point to the data

- 41:00, we're at the beginning of that, so we'll see how it goes

- 42:45, need for agile even at really small scale, with iterations: what you hope to do, the technic to do it, and what kind of data/format needed. These three points need to be flexible, and able to change
- 44:00, that's what is the most efficient, especially when no really clear idea of what we want: need to work with the data rather quickly, not a good idea to spend too much time determining the needs
- 46:15, first process is "normal", expression of the needs, then try to build it, however, we might then figure that actually the data do not exactly fit our needs ➔ small iterations, and going to the data quickly is important
- 49:30, also possible to start to work from the (technological) material, to see what is possible to do. When coming to a data standpoint, it gets blurred, now we can basically access to everything (not always easily)
- 53:45, never seen the data ➔ idea process
- 56:00, however, OSM is actually a good starting point, since it leads you to a specific (maps) domain
- 57:40, data as a starting point interesting for non-business organisations: humanitarian, citizen-driven projects, political/economic prediction
- 58:30, this is these organisations that drive the opening of data

- 01:00:30, data is usually modified to suit the system, since they try to avoid middleware
- 01:02:30, not really the good person to answer about the convergence of standards, that what makes sense, and his company usually goes for standardisation

- 00:20, value of data is subjective
- 02:30, value and accessibility of a data is not related, internal data to improve the efficiency of the logistic of an organisation would be useful for it, but not for everyone else
- 03:00, the value is also related to the notion of supply and need
- 04:00, very precise data might be interesting for people, but are they really to pay for it?

- 05:00, we can say that open data has a value for society, since it imply transparency

- 06:45, comparison with the work of a journalist makes sense, that what they more or less do in their research unit
- 09:30, XML expert is interesting, perhaps not a "full-time job", yes, interest for these data jobs (data scientists, etc.)
- 11:15, people should not be too specialised, would imply a "less-efficient" communication
- 12:30, when speaking of creation of data, artistic works is essential, there are the one who'll "think differently", which is necessary if you don't know exactly what you will do
- 13:30, need for design thinking-ish technics
- 14:00, pluri-disciplinary teams is important

- 15:30, no experience on the idea of putting APIs in front of developers

- 20:00, reliability matters, depends on how the use of the data is critic
- 20:45, however, free but less reliable data lower the value of the expensive but completely reliable ➔ if we can have weather models free, are the paid one (which can be better) still as useful? Depend on what you want to do ➔ critical data vs non-critical data
- 23:00, a lot of people wants to consult free but less reliable data, and won't pay a lot of money for the best one
- 24:40, we trust the brand

- 26:00, not sure that OD increased the innovation so far ➔ they were using partners, now you need to study the OD to see the quality, how to use it, etc. ➔ time + money
- 34:00, yes to the creation of communities around the company

- 38:00, a service, acting as a middleware, filtering and "authorising" external data to internal use ➔ verify some sources, or some data
- 44:00, even when money is not a matter, non-reliable data is interesting to start with, prototypes, beta, etc. Then ability to move to certified data.
- 45:00, also possible to make a distinction between different level of quality of data, for different publics
- 48:40, agrees with the interest for the culture in the company, especially in smaller companies, bigger ones have really specialised services and employees

J - Data summary

This document contained a summary of all the data collected during the four interviews. Between parentheses I precise which interviewee agrees (or disagrees) with each point. For instance, "Wrong format (2, 3)" says that the second and the third interviewee had problems with the formats of the data.

I highlighted several categories (explained in the Result section):

- Reliability of the data
- Preliminary work on the data
- Skills
- Value of the data
- Design process
- Characteristics of the organisation
- Description of the data
- External opening of the data
- Internal opening of the data
- Understanding of the data

Platforms (the data)

The data stored on these platforms is hard to use:

- Wrong format (2, 3)
- There is data, but its usability (interest) is not obvious (1, 2, 3)
- Poor quality (errors, outdated) (2, 3)

There is a need for a pre-treatment of the data, to be able to use it: it is raw data, not a service (1). This is why several companies provide a service on top of this data, such as OpenDataSoft or Data Publica (2, 4). Making a wrapper around the data, making it easier to work (and especially to cross it with other data) is an important step.

A collaborative community working to improve the quality of the data is unlikely to exist: corrections would not live through updates if the correction is not directly sent to the producer of the data (2). It would also require a collaborative platform that does not exist (3), and there is a problem of license and "certification" of integrity (4). However, such idea would be interesting (1).

Linked data

Description of the data is a crucial part of using it (3), makes it therefore easier to use it. Fully linked data is/would be great (1, 2), but would it answer a problem, and would it increase the complexity/scalability of it (4)? The financial world already uses a similar process with unique identifiers. This spare the industry big difficulties and is a reason of growth (3).

This has the benefits of letting people be platform agnostic, just accessing the data, navigating through it, without having to rely on a platform (1).

Internal data

There is similarities with external OD (all), since there is a lack of control on the data, and it is rare that all the services in a company use the same methods/tools, moreover, the needs being different, there is a high chance that the data is not reusable as is (4), but there might be some shared practices (1).

Like with external OD, there is work to do on the data, which needs to be usable (3), but the data being internal is interesting since you can access the producer, in order to improve your understanding of the data (1).

Opening data internally is a huge work (4), but makes it easier to open it externally (3).

Understanding of the data

There is a need for people with different qualifications, such as a person able to look and study the data the same way as a journalist (2, 4), a data specialist/scientist or a XML specialist (someone able to describe data) (3, 4), or a person with a creative background (designer, artist) to think differently (4, interviewee 2 disagree), techniques like design thinking are therefore useful (4).

The value of data is subjective (4), what matters is not as much the data in itself, but rendering it to make it usable (3). Its value is also defined by the people having a use of it (3, 4), OD, being free lowers the value of paid but better data (4).

Formats and standards

Working around formats is a hassle (1, 2), companies usually figures a way to adapt the data to suit their need, to avoid middleware (2, 4). It really depends on the situation (1, 2).

We can expect a convergence of the standards (2, 4), however with the amount of available data increasing, we might not reach a state where all formats converge (2). Instead of a convergence of the sources, we could expect to work with a third part provider (2).

Creation of an API between the developers and the data

Web developers prefers to work through REST APIs than with files (1, 2). However, it adds a layer of development (2). Generally, this really depends on the circumstances (2).

The software development process

Some consider that working with OD does not imply many changes, the work being provided by the organisation publishing the datasets (1).

Agile methodologies and iterations, while being useful in many kind of projects (and specifically research ones), are essential when working with OD (2). Indeed, when doing a project, there are three points to consider, what the software should do, the technic to do it, and the data (and its format) used; this points need to be flexible. Moreover, it is mandatory to start working with it early in the development process, since there is little to no control on the data.

While crowdsourced data is usually reliable (1), it is hard for a company to trust external data when there is no organisation to certify it (2, 4), there is therefore a need to validate it internally (2, 4), which could be done by a specialised service (4). The need for reliability depends on the criticality of the project (4). However, the value of "perfect data" is lessen by the existence of free less reliable data.

Innovation

OD facilitates the realisation of ideas and helps the birth of new ideas (1, 2), but it is unlikely to come with a big idea for a commercial service directly from data, it is more common to get an idea and then use OD during the development (3, 4). Most of the good services are the ones filling needs (3), and usually data in itself is not sufficient to come with an idea. OD is a plus; people with an idea will usually find a way to do the service anyway (3).

OD can be a source of reduction for money/time (2, 4), but with the increased amount of available data, a pre-treatment process (filtering, sectioning, etc.) is necessary, which cost time and money (4).

It is beneficial for a company to open some of its data (all):

- It can spawn new businesses hooked to the data (1, 2), which can eventually later be copied/bought (2)
- The organisation can retain public relation benefits such as improvement to the image/transparency (1, 2)
- Creation of a community (1, 4)
- In some circumstances (specifically for public organisations or monopolies), it is also a way to precede regulators which could ask for a worst outcome (3)
- It is also possible to open something to be the first one on the market, and get people picking the company's data (see Bloomberg and the BSYM) (3).

There are also risks with the opening of data:

- Breaking some law (non-correctly anonymising data) (1)
- Risk of facilitating the work of competitor/lowering the cost of entry on the market (2)
- Imply a certain transparency, therefore not possible lie/hide some stuff (we don't have access to the number of users of Twitter) (2)

Company organisation

It does not seem that the existence of OD will change a company (1, 2), but some companies are more likely to be able to work with OD than others (1, 2). For instance, a good communication if important to be able to discover datasets (1, 2), therefore a flat structure could be more efficient (1). It is essential to have access to a certain number of skills for a company, such as data specialists/engineer (2, 3, 4), data scientists (3, 4), people with the same work process than journalists (2, 4, 3 do not agree with the search for data), people able to describe/wrap the data (XML specialists) (3) and creative people, such as designers/artists (4). Moreover, these people should ideally be multidisciplinary, and the teams should not be specialised (4).

Opening data internally is an important step for innovation (1, 2), to increase the communication inside the company (2), and have a positive effect on the culture of a company (1, 4), especially in smaller ones (4).

Opening data internally can be difficult when the services are already built (4), or when some people have a monopoly on a certain knowledge, they can be reluctant to giving it back (2).

The existence of OD is interesting for a company to expand its business to fields, which are not part of its core (1, 2), but the creation of partnerships is also a possibility (2).