DARIAH Winter School in Prague

Open Data Citation for Social Sciences and Humanities

24th to 28 of October 2016

Session 9: Infrastructure & Platform

9-Infrastructure & Platform

Contrasting platforms and infrastructures as configurations for data sharing

Jean-Christophe Plantin, LSE, UK

I am the co-director for 2016-2017 of the Master program Data and Society at the LSE, where I work with students on topics such as the governance and public policies in sectors that are increasingly *data driven*. As it is in media and communications, we talk a lot about journalism and social media, but a lot of things we see apply also to the world of library, data sharing and infrastructure. The larger umbrella of my research is the *platformization of social life*, which designates services provided typically by infrastructures that are increasingly provided by digital platforms. For example, I work a lot on maps and cartography, such as the IGN in France. Since the arrival of the World Wide Web, we have seen platform-based mode of cartography that are increasingly reaching the scale, and the nature of essential service of infrastructure. This is the configuration I am working on.

In the world of archiving, we see the same tension, between, on one side, infrastructure and data archive that have been existing for a long time, and more recent web-based platforms on the other side, that present their activities as doing the same services sometimes *in addition* to infrastructure, sometimes to *replace* infrastructures.

The purpose of this talk is to compare these two entities, to "map" their relationship and to see what are the risks and benefits when it comes to data sharing for scholarship.

Data as scholarly output

Incentives:

- Rise of big data and data across disciplines
- New data sharing requirements (incentives from funding bodies)
- Diversification of materials considered as scholarly outputs: greater interest from researchers, librarian, etc. to extend the artefact of scholarly communication beyond journals (datasets, simulation, softwares, etc.)

The decentralisation of scholarly infrastructure

- Rise of the World Wide Web and the possibilities to decentralise scholarly infrastructure
- Web as technology and culture that challenged the traditional vertical and central system of scholarly infrastructures: publisher, library, archive
- Reduction of the publication cost with electronic media, scholarly productivity measure (hyperlinks as alternative to citation count), e-print movement and repository effort of the early 2000's (ArXiv.org)

Figshare

Figshare is very much a product of these two tendencies, using technical environment of the Web, adopting values and characteristics of platform, as well as positioning itself towards these new needs and incentives around data. Figshare describes itself as a "*platform where researchers can store, share and get credit for all of their research*" but the broader objective is "*improving and opening up the dissemination and discovery of scientific research*". It invites individual researchers to self-archive their outputs (datasets, graphics, presentation slides, almost anything) through personal profiles you can create, such as on Academia or Facebook. It was created in 2011 as a "pet project" from Dr Mark Hahnel, a stem cell graduate, before being a company hosted since 2012 by digital science based in London.
This is figshare.com, but Figshare also has a technological side, which is at the basis of their second target: Figshare can be deployed as a middleware service marketed as *Figshare for Institution* (e.g. with Monash University) or *Figshare for publishers* (e.g. with PLOS). It links together institution-based and publisher web portals with a custom-made data deposit and publication platform. If you are a university, a research lab, a publisher, a library, you can contract with Figshare and get custom interface, storage services, search capabilities, etc.

Figshare as a case study is a good example of a platform based-technology. We know that both platform and infrastructure rely on *principles*, they have *technical characteristics* that are different, so what happen when they are both conflated? What does that mean for scholarship? For data accessibility?

Infrastructure and platform properties

What defines an infrastructure and what defines a platform according to a series of criteria:

- Architecture
- Relation between components
- Market structure
- Focal interest
- Standardisation
- Temporality
- Scale
- Funding
- Agency of user

You can find further details in the article "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook".

Relation between components

- Infrastructure: Interoperability through standards
- Platform: Programmability within affordance, APIs

Market structure

- Infrastructure: Administratively regulated in public interest; sometimes private or public monopoly

- Platform: Private, competitive, sometimes regulated via antitrust and intellectual property

Focal interest

- Infra: Public value; essential services
- Platform: Private profit, user benefits

Standardisation

- Infrastructure: Negotiated or de facto
- Platform: Unilaterally imposed by platform

Temporality

- Infrastructure: long term sustainability, reliability
- Platform: Frequent updating for competitive environment.

=> This is a typology of these two separated entities, but it allows to see what is in between, when these two properties conflate and sometimes conflict.

ICPSR

The Inter-university Consortium for Political and Social Research (ICPSR), university of Michigan, is an infrastructure, a data archive created in 1962. They are specialized in archiving and sharing social science data, especially large-scale survey data produced and published by research institutes. They are a membership-based institution: everybody can deposit datasets on their website, but if you want to access it, you have to be part of its network of more than 800 member-institutions. They obtain data either from researchers who directly deposit datasets on the website, or they proactively acquire some datasets that their community wants to get.

Care of data through manual processing

ICPSR works as a library when it comes to data circulation, data acquisition, data processing and sharing. Researchers deposit their datasets, if it fits the appraisal criteria of the institution, all the datasets are processed the same way to fit the standard of the ICPSR. Data processors are the people who make sure deposited datasets are “clean enough” to fit the criteria of the institution. The reconstitution of the "pipeline" for data processing includes the following steps:

- Deposit of dataset
- Dispatch
- Repair
- Contact with the PI (principal investigator)
- Prepare
- Verify
- Publish

Every submitted dataset go through this pipeline. The two specific actions that I want to stress here are what I call “repair” and “prepare” because it is the action of data processors that really reflects this care of data. *The institution consider by default that datasets are not* *perfect, are "broken" or contain mistakes and problems, so they hire dedicated people who putting their expertise and work time to take care of it before publication*.

- "Repair": data processors rely on different scripts and softwares to go through huge SPSS files, and to flag some mistakes, inconsistent code, some missing documents, etc.
- "Prepare": data processors make sure data is presented in a way that fits the ICPSR standards, to make sure every document is presented in a specific way, using templates, make that metadata fit with the catalog of the institution.

Figshare: Automatic data provision

Of course, it contrasts with Figshare, which has a very different model. On Figshare.com, the website where you have a profile, you can just drop your dataset. There is a minimal curation, there is no added value, no labour. Data circulation is also different:

- No processing, self-deposit
- Centrality of the API (Application Programming Interface)
  - To connect web-based actors with the scholarly world (ex. reference platform like Zotero and Mendeley, repositories like Github, academic or institution libraries, cloud storage, online scholarly journals). All is mediated through their technology, their API.
  - To connect with institutions and publishers (PLOS and Monash University), this is how they design custom search capacity or custom portals.

To understand the way APIs mediate between different actors, you can read "Code as a research object". Figshare, Mozilla Science Lab and Github got together in this project and designed a Firefox browser extension that generates a DOI for datasets, code are deposited on Github and can be released on Figshare. Different platforms connect together because they talk the same language. Here, the API is central, and constitute a brokerage point between Github and Figshare, using their APIs and developing systems so that any Github repository can be processed as a package. We see here how technically a *web based data model circulation* is applied to the data sharing and data reference.

Consequences on scholarship

Relations between these two entities for scholarship:

- Infrastructures: for example, ICPSR developed an expertise over more than 50 years, network of members, reputation, recognized standards and quality, labor intensive, slow but considered as maintenance work, so there is an high turnover, so you need to train again new people, so it costs time and money for an institution.
  - Manual processing for specific type of data, but what about big heterogeneous data?
  - *Path dependence* and *reverse salient* (Hughes 1983, historian of technical systems who studied electrification in the US): when you look at how the ICPSR works and how it developed an expertise, they are very good at what they are doing but there is a path dependency towards one specific type of datasets. They became extremely good at it, they are highly specialized

mostly on survey data but as a result, they have tremendously narrowed down the amount of datasets they can accommodate. If you want to work with other kind of data, you have to redesign the whole pipeline.

- Platforms:
    - Figshare presents itself as having a strong commitment to open data and open science. Mark Hahnel is involved in lots of scientific open data events. He is very active in the community. Figshare released a report on the state of open data. They also have a clever way to contact with libraries, they integrate themselves with the existing standards, digital preservation network, citation survey such as data science. They are doing it in a great way if we have open access as goal. But there is no mean, technically or philosophically, to make sure this commitment is going to stay for the long run. We have a lot of examples of platforms who changed tremendously their data access with regards to a change in their business model, cf #DeleteAcademiaEdu because Academia.edu added a new feature going towards freemium model but we don't know what is going to happen afterwards.
    - *Splintering infrastructures* (Graham, Marvin, 2001). They use this term for urbanism in a huge study on cities, showing that with the increase of what they characterize as neoliberalism, cities that usually had a provision of essential services with infrastructure are being replaced by what they call "*networked premium spaces*". It is the idea that instead of having essential services for everybody, we have pockets of priorities where people are going to have customized access to the services. So the infrastructures are “splintered.”

Conclusion

=> More heterogeneous data
=> More incentives to deposit
=> Two candidates to accommodate these data: infrastructures and platforms

- Infrastructures developed a specific expertise following high standardisation but making it hard to accommodate a wide range of data.
- Platforms jump right in in that situation, using the flexibility and plasticity of its structure to organise data sharing, direct deposit, API. But with different configuration come different forms of care through data processing (absence of care through automatic provision).

Platform can commit to openness but there is no way to make sure there is commitment on the long run. Platforms are part of a larger decentralization of traditional scholarly infrastructure, a risk that can emphasize the tendency of the splintering of infrastructure by showing how institutions can do more with less.

Contact

Jean-Christophe Plantin is Assistant Professor at the London School of Economics and Political Science, department of Media & Communications. He investigates the civic use of mapping platforms, the collaborative challenges in big data science, and the evolution of knowledge infrastructures. His research was funded by the Alfred P. Sloan Foundation, the Gordon and Betty Moore Foundation, the European Regional Development Fund, and the University of Michigan MCubed Program. His work was published in New Media & Society, Media, Culture & Society, and the International Journal of Communication.

Twitter: @JCPlantin
Email: j.plantin1@lse.ac.uk

Huma-Num infrastructure

Nicolas Larrousse, Huma-Num, France

We prepared this presentation with Jean-Christophe Plantin with the idea to show a (relatively) conservative infrastructure compared to those new platforms which are really dynamics and different in their goals. I will focus on our own sort of Figshare in Huma-Num which is Nakala in order to show the differences of approach.

During the last decade, we saw, as Jean-Christophe Plantin showed, that Humanities are really in a *digital turn* and regular researchers deal now with digital data. It means that there is a need for tools, softwares, mediation between raw data and researchers. They also need to appropriate data, they need to shape data to do something useful for their research and this is a long process, from the beginning to the end. Between, there is a huge cycle. And now that we have more and more data, it is getting a problem. It means that researchers can't work only with their personal computers anymore. They need to have some tools able to deal with the amount of data produced. At Huma-Num, we also see that there is a need for more sophisticated tools, the classical Filemaker is no longer sufficient. For instance, they use Geographical Information System and you need to be aware to use it. They also need to preserve data which is, in my opinion, a great problem for the future of research.

But what is an infrastructure today? Is it a data center?
In the definition of the European Community, it is a whole set of things. Of course, you need to have some computers and a data center somewhere but you also need to develop expertise and a network of people.
=> An infrastructure is no longer computers somewhere. Above all, you deal with data. So you need to show it, to share it, to disseminate it and to preserve it.

Huma-Num is a French infrastructure for Humanities. In order to address this kind of question, we support groups of people with expertise (we call them consortia) but we also provide virtual machines, softwares, hosting, preservation, storage, etc. We organized it as a sort of onion with at the center the "users" (researchers, network of researchers, projects). We also create consortia, that means that we fund group of people who share the same interest about scientific objects, not necessarily coming from the same discipline, and so they can work together. We expect from them to build expertise, good practices, tools, standards etc. that we can share with other communities. With this process, we try here to build a *virtuous circle* which is really the center of our project. And beside that, we provide some core services, machines, virtual machines, softwares which are generic services, but we also provide specific services centered on data: that our way to fit to the technical needs of research projects. At Huma-Num we want to offer original services to the community for data processing.
The last part of the onion is that we need to have exchange with other people doing the same thing, especially in Europe. European Infrastructures for Humanities are on the way (e.g. DARIAH) and Huma-Num intends to be a sort of hub to Europe for the French community in Humanities. The main idea is to valorise what the French community is doing

(tools, expertise, data). In return, we can get information about the way communities are organized abroad and what they are doing, so we can situate our actions and reinforce networks of expertise.

Nowadays, Huma-Num funds about 10 consortia in very different disciplinary fields (Musica: music encoding initiative, etc.). The idea is to have people form different structures to work together, which is not always easy in France. For example, we have the 3D consortium, which is not disciplinary, it is a group of people interested on 3D in general for their work, it can be archeologists or people coming from geography. There is another consortium, Archipolis, on Political Science which concerne is long term preservation of surveys. They invent their activity in fact, for example new set of metadata to describe a common object (3D or surveys), because standards need to be adapted in order to address new needs.

The second pillar of Huma-Num is technological services: along big storage, you have softwares and hosting websites, virtual machines. The infrastructure is hosted in an already existing data center, we didn't want to build a new one for Humanities in France, it is a total non-sense for us. So we pay to be in a huge data facility specialised in physics and so we can provide a lot of storage, garanty power and network availability, etc.

The originality of Huma-Num is about data services - it could be an issue in Humanities because researchers generally store data on their computer and it can get lost easily (e.g. computer crash, end of fundings, retirement, etc.). There are a lot of *nice* ways to lose data. For instance, if you have a project funded for three years and you have a beautiful website done by a private contractor, a postdoc or an engineer and then the project stops to be funded. As your data is in the website, it is lost because nobody is going to maintain it. Besides the technology gets obsolete rapidly nowadays. So, after a while, there is no way to access data and to cite it.
So we try to provide a set of tools in order to preserve data and the preservation for us is to be able to share data, to disseminate data, to inform people that data exist. We also provide long-term preservation service, which is more specialized: for this, we rely on an existing specialized data center. Our added value for this data center is to provide them with new type of data and metadata, it helps them to improve their way of dealing with the long term preservation service.

NAKALA & ISIDORE

Since many scientific data producers do not have the digital infrastructure to provide persistent and interoperable access to their data, Huma-Num has implemented a tool to expose and share research data called “NAKALA”.
NAKALA provides mainly three types of services:

- A PID (Persistent IDentifier) to data and metadata
- Permanent data access
- An exposition of metadata through a Triple Store and OAI-PMH

NAKALA is a simple repository for sharing resources:

- The main API to access data is the Triple Store
- You can cite your Data and your MetaData

- Data and MetaData are immediately available

But if you wish to show your data, you need another application, not provided by NAKALA:

- A search Engine
- Tools for visualization

We decided to develop NAKALA in order to address the issue of data being lost in a website. NAKALA is a repository with the idea of separating the place where data is and the place where you show it. As Jean-Christophe Plantin mentioned, there is nowadays a competitive war between technologies. When you choose a technology, after a few years it will no longer be trendy or available. This is why data should be somewhere else to be able to share and show it on the mid-term.

So NAKALA is this repository and then you will have to build something upon it to show it. In Huma-Num we use Omeka to do so. It is not really flashy but it works very well. Then we connect NAKALA and OMEKA. If one day you want to get rid of OMEKA, it is not a problem, your data is still untouched in NAKALA. You can also use hypotheses.org for example to show data from NAKALA.

ISIDORE is the place where you harvest a lot of repositories, including NAKALA but not only. ISIDORE harvest about 4000 sources and show 4 Millions records. Then there is a chain of treatment to enrich, classify and link all these metadata. Isidore only deals with metadata, never with data, it attributes a handle to be able to cite your data and there is a huge work, a sort of *automatic curation of metadata* to enrich, to classify and to put it in Linked Open Data.
=> Every word or term used in Isidore is linked to the LOD by using Semantic Web Technologies.

So, you put your data in NAKALA, it is safe, Huma-Num takes care of that. Then you can advise people your data are here by connecting Isidore. Isidore will harvest metadata, classify them and disseminate them. ISIDORE can be viewed as a specialised search engine, but Google is also really fan of Isidore semantic classification. We also have a Triple Store In ISIDORE as well as in NAKALA. NAKALA is totally built on Semantic Web technologies, there is no relational database.

Then, when your dataset is complete or finished, you can preserve it on very long term, for example now we deal with huge set of digitized manuscripts. We work with the French National Computing Center for Higher Education (CINES) which has the official mandate of long term preservation for scientific data - at the beginning it was dedicated to thesis in digital format). So, in France, data produced by researchers from public money are supposed to be, one day, on the National Archive. Long-term preservation is a specific process, based on the archive field. Basically we put it on special devices in the CINES and they take care of it, they take the responsibility of data by making copies and they try to re-read data every month; if the format is obsolete, they will convert it and this is a huge responsibility. So they make sure that in 20 years we can re-read and understand this data by adding metadata and context information. There is a technological part, but also an archivist one.

An example of good tool for today's interoperability: semantic web technologies
Even if it not easy to use for people, it is perfect for machines as it works greatly to exchange data and to link data to other repositories too. We provide tools for hosting Triple Store content, the basis of semantic web. The idea of NAKALA and ISIDORE is that if you are a researcher you can put your data in NAKALA, or to make your repository harvested by ISIDORE, then you are automatically in the graph of semantic web, even if you don't know anything about semantic web.

We have a lot of links with French repositories like data.bnf.fr, which the National Library and it is remarkable to see that the National Library publishes this kind of data in RDF since 3 years whereas they are working with these technologies for about 10 years now. This is because in fact, you need a lot of curation work to permit people to publish and to maintain this kind open data repository.
There is also the DBpedia project, which is supported by the Ministry of Culture for the French part of Wikipedia. We also use geonames, lexvo and other repositories.
Nakala and Isidore provide a SPARQL EndPoint: for NAKALA it is considered as the API to access the metadata associated to NAKALA's TripleStore.

NAKALA is in fact associating different bricks, which you can replace. For example if you want to get rid of this TripeStore which is today very trendy, you can replace it; if you want to change storage, it’s possible; if you don't want to use an other OAI-PMH software, you can change it etc.
NAKALA is organized around a TripleStore, we don't have any database, so we will be able to switch to another technology in the future.

So we deal with data, any kind of data: it can be code, archeological data in a zip file, voice recording. Still NAKALA provides with some tools to suggest good format, for example when you add a picture, you have a tools that checks formats. To decide what is *right or wrong*, we rely on the work done by the CINES, which maintains a list a recommended formats. You can add Filemaker projects, but we suggest that it could be better to prefer another more reliable on long-term format. The only requirement is to give at least four elements of metadata which are title, author, date, type. Deposited data are securely stored and a PID is attributed to it. Figshare uses DOIs, we give handle. It is the same old technology. For metadata, you can access it through the TripleStore. Everything in NAKALA is organised around TripleStore. So as soon as you add your data to NAKALA, it is accessible.
The curation is made when you ask for an access to NAKALA: Huma-Num does an evaluation of scientific goals of the projects as well as the future of data (e.g. openness, licence etc.).

Data deposit in NAKALA

There is a web interface, not as sexy as Figshare, still it is easy for researchers to add for example 20 or even 50 videos and to share it. If you have more expertise and more data, for example an archeological project in Egypt uploaded 120 000 pictures from the walls of Karnak, you can use batch processing, add metadata in XML, make a packet and process it

into NAKALA. PIDs are attributed to data and metadata: the “handle” technology for PID (used also for DOI) gives two possibilities to access data: using a NAKALA URL or a regular handle URL.
For the access, right now, we don't provide any specific API because the TripleStore is supposed to be the API, but it might evolve in the future.
To get back to the main topic of the talk, we can say that NAKALA is not really a platform, but more an infrastructure. For instance, it doesn't provide any visualization tool, the idea is to have a simple repository. Huma-Num provides on the long term the maintenance of the handles system because the citation is really useful for data and metadata. And we also hope to provide with usage statistics: a good measure of the infrastructure use.

To display your data, you will need something else built above NAKALA, for example you can use Wordpress provided by hypotheses.org. Huma-Num provides a CMS OMEKA to display easily your set of data plugged with NAKALA that we called NAKALONA. For example, a French research center in Kenya made a 40 years press archive and added it to NAKALA. We had several exchanges with the persons in charge of this work and we decided to use a batch processing because there are more than 10 0000 pages. They want to share it at large, to show it, so we used NAKALONA and within minutes it was possible to show it or to search with metadata (see https://ifrapressarch.nakalona.fr).

Others can develop their own interface, etc. And Isidore can be considered as an interface. There is no limitation in term of size of data. We use a distributed technology called Active Circle that allow to aggregate parts and make an abstraction of the physical storage, so the size of the storage is not a problem. Anyway, we now think about providing more curation, more advices about good practices, helping people even to prepare data to go to NAKALA, but there is an important human cost. The issue is not to have only good tools or infrastructure, it is also to make people use them.

=> The problem is that there is no reward, no incentive for properly sharing data (with sufficient metadata). The only incentive is to be cited so today it is like a loss of time for a career. For Figshare, it is a question they have. For the self-deposit website, the main difficulty is to get people adding metadata in addition to just dropping files with no description. Sometimes, if they see that a dataset has an important download rate, they contact the researcher depositor and ask him to improve it, but it is not systematic, it is a kind of *download driven metadata*.
For publication, in ISIDORE, we can try to find if there is a PID related to a dataset and add a link to metadata. Like this, we try to build links between datasets and publications in the sense of RDF and semantic web graph.

Incentives for open data
In France, when you apply in humanities through the National Research Agency (ANR), you can declare that you are in touch with an infrastructure, but it is not yet mandatory. However with DMPs in H2020 projects, things are changing a little bit. Funders are aware that data cost a lot of money, so they need to preserve it and to share it. In France, it is just the beginning.

Contact

Nicolas Larrousse is head of the long-term archiving department at Huma-Num, a French infrastructure which aims to provide services to researchers in social sciences and humanities. He is particularly focused on interoperability and is involved in European infrastructures and projects. Huma-Num is promoting collaboration and providing services to manage, enrich and expose research data through a wide network of partners and consortia. Huma-Num is the National Coordinating institution of DARIAH European infrastructure for France and is involved in H2020 European projects.
Huma-Num: http://www.huma-num.fr/
Twitter: @Huma_Num
Email: Nicolas.Larrousse@huma-num.fr