Harvesting and Aggregation of Research Activity Data

Cecilia Loureiro-Koechlin
Project Analyst
SERS, Bodleian Libraries,
University of Oxford
Tel: 01865 280028
E-mail: cecilia.loureiro-koechlin@bodleian.ox.ac.uk

Data harvesting and data aggregation are terms we come across very often nowadays. They are normally used in online (web, internet) contexts[i], and refer to the manipulation of large volumes of online data to facilitate their access, analysis and preservation as well as to improve their visibility. With the amount of online data growing exponentially and lacking structure, finding information on the web becomes harder. In this changing context harvesting and aggregation become crucial processes to incorporate some order and facilitate access particularly to information professionals. But, do you know exactly how data harvesting and data aggregation are done? Do you know what kinds of transformations data go through and how these transformations can help us?

In this paper I will present an account of data harvesting and data aggregation (with no code!), stating the reasons why such processes are needed. This account is based on the work we are doing at the Systems and e-Research Service (SERS) at the Bodleian Libraries, University of Oxford. We are working on an entity registry which harvests, processes, stores and re-uses information about research taking place at the University of Oxford. A pilot of the entity registry was developed as part of the Building the Research Information Infrastructure (BRII) project[ii]. During the life of the project a sub-set of data were harvested as a proof of concept. Having successfully achieved the objectives of BRII, we are now expanding the registry to collect information about all academic areas across the university.

The kind of data the entity registry collects is called Research Activity Data (RAD) and comprise descriptions of researchers (names, affiliations, research interests), projects (names, participants, funders), and outcomes (publications). RAD come in various shapes, for example long descriptions as in researchers' biographies, or lists of words or phrases like research interests. The entity registry harvests publicly available RAD from internal and external sources. (It does not collect sensitive data such as financial data and information which is part of confidentiality agreements.) Within the University of Oxford RAD are displayed in hundreds of disconnected and disparate online sources such as departmental, project and researchers' websites as well as stored in departmental databases and spreadsheets. RAD about Oxford are also found in external sites such as research council websites, online journals and databases.

Once stored in the registry RAD can be re-used by users through the Application Programming Interfaces (APIs) developed by BRII. The BRII project also developed the Blue Pages[iii], a directory of research expertise, a kind of search engine which accesses RAD in the registry to produce Researcher and Research Activity profiles.

Research Activity Data are essential at academic, administrative and strategic levels. Information about researchers, current and past research, publications and collaborations is frequently solicited by staff who are involved in or support research at the university. When the task is about finding specific, small sets of data the solution is usually to look for that information in a particular source (e.g., a website). However when the task involves the collection and combination of data from different sources the solution can be harder if not impossible to find. I will give you two examples:

- Compile a list of recent publications on public sector management, eGovernance and other related areas authored or co-authored by at least one Oxford researcher. To create such a list you will need to access several sources of information such as online journals, databases and researchers' profiles. In departmental websites, researchers list their publications which usually include journal papers, books, book chapters, conference papers, etc. Some researchers have profiles in more than one site as they are affiliated to more than one academic unit, and the profiles and lists are slightly different between them as they are written for different audiences. You may also find a new range of keywords which are related to the main two (public sector management and eGovernance) but which are equally used by researchers and journals to classify their work.

- Write a report about the academics and groups who are doing research about China and India, or the people who are collaborating with the University of Bologna in Italy, or with Microsoft. For the report you will need to include researchers' names, current projects and publications. For this you may need to search across a wide variety research fields and websites. Researchers may be working on the history of China, on economic policies in India, clinical trials with researchers in Bologna or computer experiments with Microsoft staff. The variety is such that you cannot limit your search to a small number of research fields.

These kinds of questions are not uncommon in Oxford but are laborious and time consuming to answer. Searching across departmental websites one by one or using Google will not guarantee that you get all the relevant information you need in time. However, if you create a resource like the registry where you can have all these sources of data, not only in the same place but interconnected and organised, you can save effort and time as well as make sure you get most of the relevant information you need.

How harvesting and aggregation of RAD are done?

I asked Anusha Ranganathan, my colleague and a software engineer at SERS to explain the design of these processes and I have summarised (and decoded) her explanation here. The design follows the ideas developed by the ReSIST project[iv] at the University of Southampton.

First, harvesting and aggregation have become almost synonymous, as most harvesting processes involve aggregation and most aggregations need harvesting of data. Second, there are three main steps that are followed to feed the registry with data. These steps are:

1) **Extraction of data from original sources**. This is the actual harvest. We make a copy of these data, as we get them, in the registry. Data formats vary but the most common are RDF (Resource Description Framework), XML (Extensible Markup Language), AtomFeeds and RSS (Really Simple Syndication) which are machine-readable and human-readable HTML[v] (HyperText Markup Language). We also get data from databases and spreadsheets.
2) **Conversion of data**. This is the core of the aggregation process. We convert non-RDF sources into RDF format allowing aggregation. RDF is the format in which all data in the registry are stored.
    a. Converting data to RDF format is a straightforward task if the source is in a machine-readable format. Databases are also easy to convert as their data are well structured. However HTML sources and spreadsheets can be extremely difficult and time consuming. HTML requires screen scraping (manual labour) and spreadsheets are normally not consistent. In general, whenever possible, we try to avoid using static files (spreadsheets, word documents, etc.) as they are difficult to update.
    b. When the original source is in RDF format there is no conversion required.
    c. RDF is a standard that uses controlled vocabularies (taxonomies and ontologies) to label and classify sets of data so they can easily be identified by machines. (See figure 1.) In the registry we are using ontologies such as ‘Friend of a Friend’ which is used to describe people, their activities and their relations to other people and objects. For example, name, research interests and X works with Y. We have also developed an ontology, Academic Research Project Funding Ontology (ARPFO)[vi], to describe funders and their relationships.
    d. After conversion, data in the registry become entities with attributes and the relationships between them. For example, entities are people, research projects, academic units, funders and publications. Attributes are the name of a person and the year of a publication. Relationships are researcher A is affiliated to academic unit Z, collaborates with researchers B and C and has written papers M and N.

e. The relationships identified at this point are the relationships contained in the original sources.

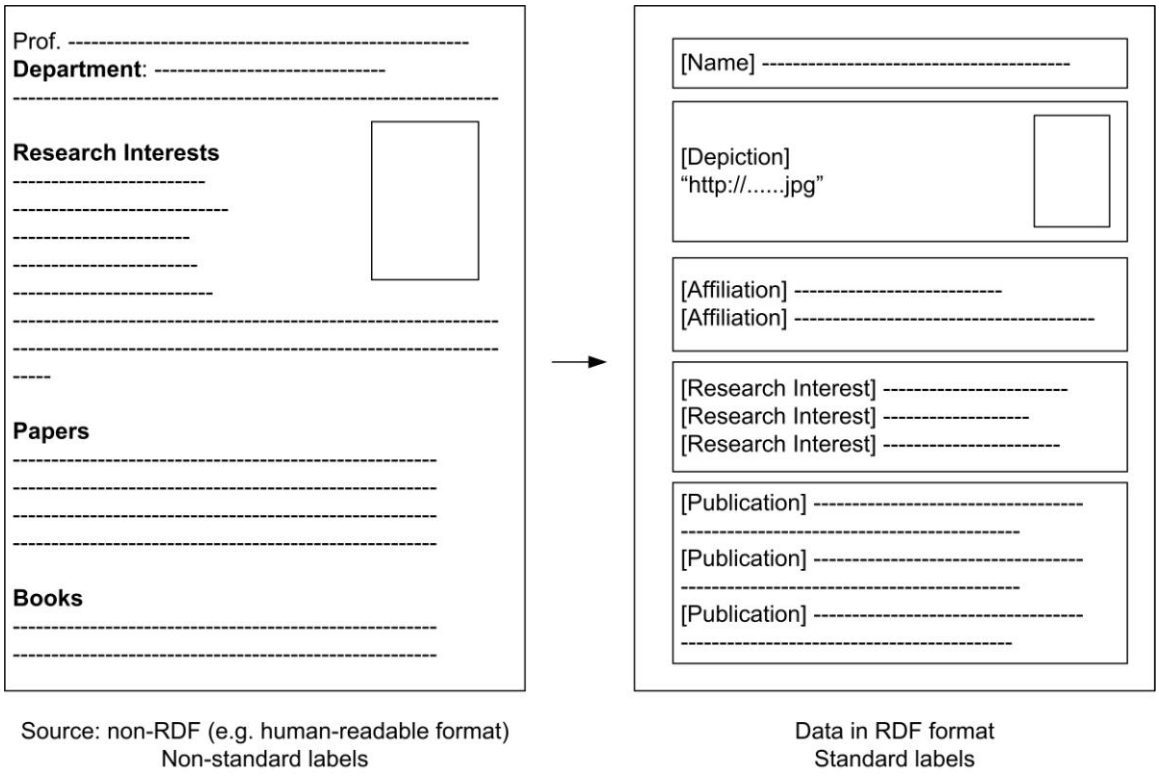

Figure 1. Conversion of data

3) **Uncovering of hidden connections**. When data are in RDF format we look for further relationships. Relationships which were not explicit in the original sources but which are possible to identify now that data are labelled. This stage deals particularly with connecting data from different sources. For example if we get the name Prof John Philip Smith in sources 1 and 2 we can establish with some level of certainty that these two sources refer to the same person. Therefore we can connect the data in these two sources. There are, however, other cases which are not so straightforward, where names are similar but we cannot be sure they belong to the same person. For example if we get Prof John Smith's biography in source 1, Prof J. P. Smith listed as Principal Investigator on a project in source 2 and John P. Smith as author in a publication in source 3. For cases like these we have developed a 'same-as' process (See figure 2.)
    a. 'Same-as' has a set of rules which use information such as people's first name and surname, researchers' affiliation and email. Depending on the availability of information and whether the sets of data match, 'same-as' will determine if two or more records belong to the same person or not. If the records belong to the same person 'same-as' will merge the records (number 3 in figure 2). If the information available is

not enough to do the matching, or if the data do not match, ‘same-as’ will keep the records separately (number 2 in figure 2).

b. The ‘same-as’ process focuses on ‘people’ entities. Projects, publications, funders and academic units, usually have fixed (or standard) names which are used consistently across sources. However, names of people are frequently written in different ways, depending on the contexts.

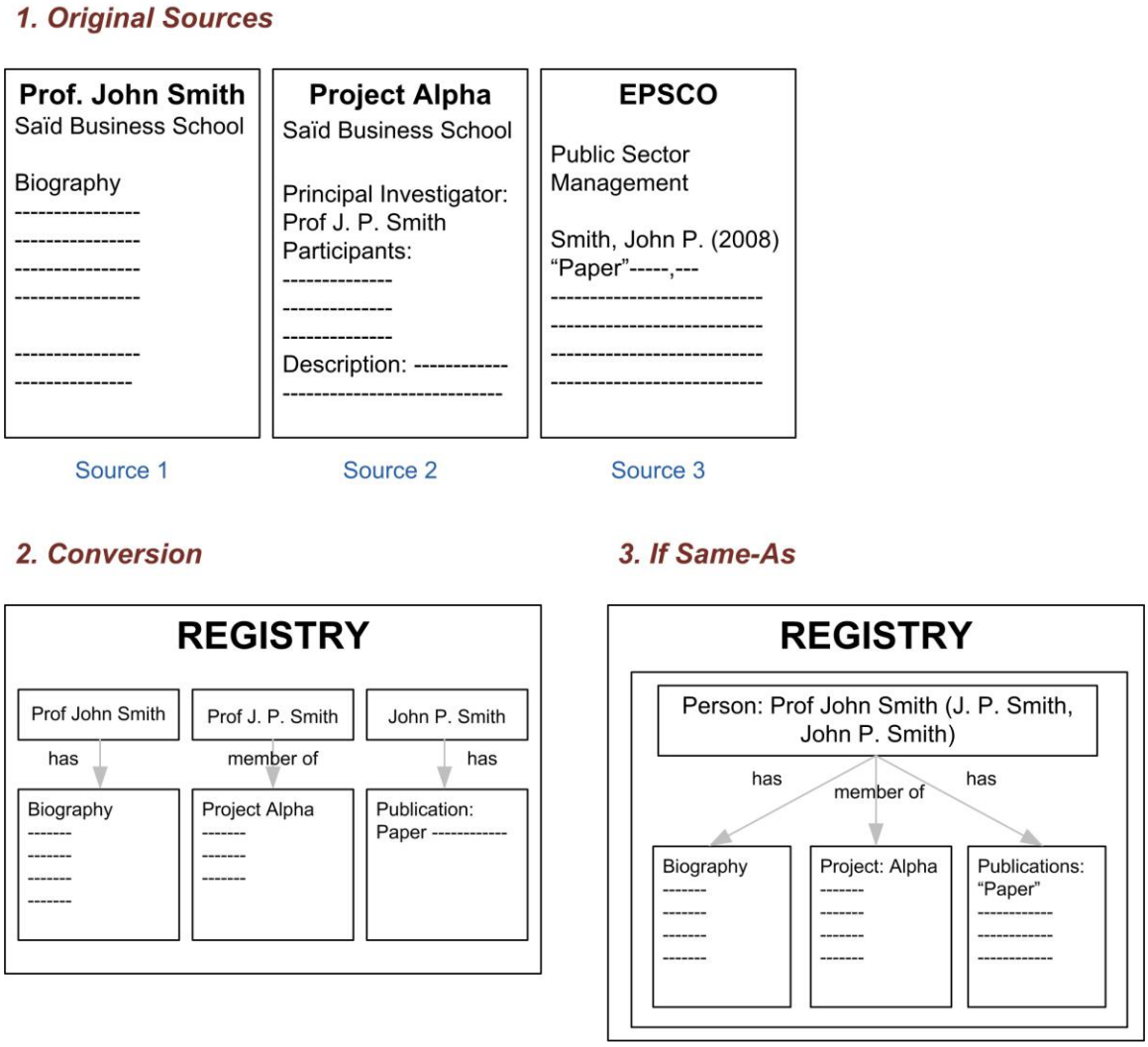

Figure 2. The ‘same-as’ process.

Third, there is a need for regular updates of data so the registry can be kept up-to-date with changes in the sources. Depending on the kinds of sources, data can be ‘pushed’ to us whenever a change is made in the original source, or data can be extracted again according to a set schedule. In both cases the steps above are repeated with little changes.

After aggregation, data in the registry become more accessible allowing tasks like the described in the two examples above to be carried out more efficiently. Other complex tasks are also possible. For example the identification of networks of researchers interconnected by their collaborations on projects, publications or research interests, or the identification of subject networks and the relationships between those subjects.

Data quality control

Due to the disconnected and heterogeneous nature of sources of RAD, and the complexity of the data aggregation process there is scope for error. Therefore data quality control is essential to ensure the reliability and accuracy of the information held in the registry.

At SERS we are outlining a system for data quality control. Part of the system will involve users flagging up errors to us from the Blue Pages, and a workflow to track the reports and corrections done to the data. We foresee two kinds of errors:

1. Errors in the content of data in original sources. That is, data that was harvested with errors. Typical examples are typos and outdated information. As the error is located in the original source, our task is to notify the person in charge of the source so they can verify the report and correct the data if necessary. Once data is corrected it will be reflected in the registry the next time it is updated.
2. Errors in the matching of records, for example the biography, photo and research interests are correct but the publications do not belong to the researcher. This can happen when the 'same-as' process merges records of two different people with similar names. One solution would be to split the record in two, one connected to the biography, photo and research interests and the second connected to the publications.

The future

Work on the entity registry is still ongoing. We are aiming at comprehensive information coverage by harvesting data from as many sources as possible and covering a wide variety of research fields. Feedback from stakeholders has been positive, reassuring us that a service like the entity registry is definitely needed to aid administrative processes, improve research visibility, increase research impact, and boost collaboration and funding.

---

[i] Aggregated data can also be found in databases, repositories and data warehouses.

[ii] The BRII project was funded by the JISC and was a collaboration between the Bodleian Libraries and the Medical Sciences division. It ran from September 2008 to March 2010. BRII website http://brii.bodleian.ox.ac.uk and blog http://brii-oxford.blogspot.com

[iii] See http://brii-oxford.blogspot.com/2010/03/blue-pages-video-clip.html and C. Loureiro-Koechlin, 'Uncovering User Perceptions of Research Activity Data' *Ariadne*, 2010, Issue 62, January 2010 (http://www.ariadne.ac.uk/issue62/loureiroKoechlin/)

[iv] The ReSIST project built the RKB Explorer (www.rkbexplorer.com) a semantic web application
[v] Machine readable data is data that computers can 'understand'. Human readable data is data that can be naturally read by humans. Unlike humans, machines need extra information to decide for example what piece of data corresponds to a name or a title. That is why we need some sort of data tagging. Humans can usually infer this from context and previous knowledge.
[vi] The ARPFO ontology developed by BRII is stored at http://vocab.ox.ac.uk/