Metadata towards an e-research cyberinfrastructure. The case of French PhD theses

Jacques Ducloy, Jean-Paul Ducasse, Muriel Foulonneau, Luc Grivel, Diane Le Henaff, Yann Nicolas

Metadata towards an e-research cyberinfrastructure

The case of French PhD theses

Jacques Ducloy
INIST / CNRS
jacques.ducloy@inist.fr

ARTIST
artist@inist.fr

Jean-Paul Ducasse
Université Lyon 2
ducasse@mail.univ-lyon2.fr

Muriel Foulonneau
University of Illinois at Urbana-Champaign
mfoulonn@uiuc.edu

Luc Grivel
Université Paris1
luc.grivel@univ-paris1.fr

Diane Le Hénaff
INRA - Centre de Versailles
lehenaff@versailles.inra.fr

Yann Nicolas
ABES
nicolas@abes.fr

Abstract:
This paper analyses metadata practices and needs in the French research community. It focuses on PhD theses whose life-cycle is totally controlled by the academic institutions. It uses information treatments dealing with setting up research policy as samples for an e-research orientation. Several case-studies illustrate the fundamental role of various repositories containing affiliations, authorities or linguistic items. ARTIST, the collective author of this paper, is introduced.

Keywords:
Metadata, vocabularies, research policy, theses.

1. Introduction

This paper is the result of a collaborative work and was written by a networked team of people, engineers or librarians, working in different organisations, in the framework of ARTIST[1] (Appropriation par la Recherche des Technologies de l'Information Scientifique et Technique) project. Our first experience was based on various contributions on a

[1] < http://artist.inist.fr/ >

terminological forum, about a translation[2] of "*What Is a Digital Library anyway, anymore*", a paper written by Carl Lagoze, and whose subject deals with the deep structure of a Digital Library[9]. This paper is a new cooperative experience which would like to analyse how metadata could help the French academic community in building a federative Digital Library.

The annual issue of the "Academic Ranking of World Universities" [6] is causing discomfort in those in charge of setting up research policies. Improving the quality of metadata items such as affiliations is now considered as a key issue for improving the visibility of universities. The researchers themselves are now permanently looking at impact factor. The "publish or perish" notion is now used as a strong incentive for author self-archiving in institutional repositories [4].

Academic librarian and research communities begin to feel that metadata are not only useful for information retrieval but could play a more strategic function. This new way of viewing is perhaps a first step towards a more global analysis about the role of scholarly publishing in what is called "cyberinfrastructure for e-science or e-research" [10].

In this context, this paper will explore how metadata could be used in some activities dealing with research policy in a francophone[3] environment. We have chosen to focus on PhD theses because their life-cycle is fully controlled by academic institutions; but a large part of the discussion could be applied to all items of scholarly publishing.

We will show that a precise research policy requires sophisticated metadata. In an open archiving framework, the most popular among technical solutions, such as DSpace [12], or Eprints[4], do not require a depositor to provide strongly structured metadata. Most requirements are limited to a basic set of Dublin Core elements in order to be easily harvested. PhD theses are naturally concerned by this goal of improving visibility [5]. We will show that their initial life-cycle requires that metadata should not be merely descriptive but should include some management elements. Indeed, most of the time and more specifically in a French context, several institutions or organisations are concerned and must cooperate.

As for all published items of research, theses metadata must be usable in any portal (national, international, thematic…) that could increase their visibility. They should also be easily handled by informetric tools in order to be picked out in a scientific or strategic watch or for research policy oriented studies. At this level we will show that a key issue is the handling of vocabularies and affiliations.

In the first part of this paper, we will start by introducing the francophone environment. Then we will present several structuring initiatives dealing with PhD thesis production, union catalogues and institutional archives. Finally, we will discuss three case-studies showing various aspects of metadata and vocabularies.

2. Digital libraries for e-research: an overview of European, francophone and French contexts

Francophone research institutions must position themselves in relation to a variety of existing national and international frameworks.

[2] < http://artist.inist.fr/article.php3?id_article=245 >
[3] From the French speaking area
[4] < http://www.eprints.org/ >

They take part in international standardisation initiatives. They have to take into account the evolution of standards and practices in the United States and worldwide. Additionally, they are part of both linguistic and regional networks. France and Belgium for instance are part of both Europe and the francophone area (Francophony). Algeria, Morocco and Tunisia are part of Francophony as well as of the Arabic language community.
As a result, francophone research actors must coordinate with a number of initiatives in multiple areas of cooperation. The metadata strategies adopted for scholarly publishing must ensure interoperability of francophone scholarly material in all those networks. They must reflect very diverse administrative situations in the different countries as well as in the regional and international network infrastructures.

2.1 International context of e-research

The open access movement and the Open Archives Initiative have encouraged research institutions to make available theses and dissertations on the Web. On the technical side, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[5] makes it possible to share and exchange metadata about scholarly material. This has allowed the creation of an open framework for publishing theses and dissertations. They are integrated into open repositories and shared in larger networks. In France, this led to the creation of the Centre for Direct Scholarly Communication[6] (CCSD), a major initiative aiming at reengineering the processes of scholarly communication, as illustrated below (section 3.2).

In the United States, efforts to create an open digital library framework in the scope of the Digital Library Initiative DLI-I and DLI-II funded by the National Science Foundation have led to such major projects as the National Science Digital Library[7]. NSDL has contributed to the promotion of standards and the development of services based on an open architecture for digital libraries. The Networked Digital Library of Theses and Dissertations (NDLTD)[8] [13] has developed an infrastructure, including processes and workflow for electronic publishing of theses and dissertations. It has raised IPR issues related to ETD (electronic theses and dissertations) publishing. It has also improved repositories technical interoperability by encouraging the use of OAI-PMH and SRU servers. Finally it has improved metadata-related interoperability by adopting the ETD metadata set (ETDMS) [4] developed as a Dublin Core application profile. ETDMS is notably used in the Cyberthèses project (francophone portal for ETD) further described in section 3.1. Alternative metadata formats such as MARC and MODS (Metadata Object Description Schema maintained by the Library of Congress)[9] are also used. The Metadata Working Group of the Texas Digital Library has developed a descriptive application profile for electronic theses and dissertations in MODS[10]. Finally, a number of libraries embed descriptive metadata in METS wrappers (e.g. The Florida Center for Library Automation[11], or Uppsala University[12]).

The ARTIST project, collective author of the present article, is notably in charge of tracking information on the multiplicity of existing metadata initiatives and their evolution in order to ensure that French and francophone actors benefit from those initiatives. It aims to better coordinate the standardisation efforts in the different networks.

[5] <http://openarchives.org>
[6] <http://www.ccsd.cnrs.fr/accueil.php3 ?lang=en >
[7] < http://www.nsdl.org >
[8] <http://www.ndltd.org/index.en.html >
[9] < http://www.loc.gov/standards/mods/ >
[10] <http://www.tdl.org/projects/metadata/tdlappprofile.pdf>
[11] <http://www.fcla.edu/dlini/etd.html >
[12] < http://publications.uu.se/theses/index.xsql?lang=en>

2.2 The European context

The European IST (Information Society Technologies) program, like the DLI programs in the US, has focused on the research dimension of information technologies to create an open digital library framework. Several projects, such as the Open Archives Forum[13][11], have been funded by the European Commission to raise awareness of national players and to investigate the technology issues related to scholarly communication.

The standardisation of the European Research Systems is also supported by the Commission. For instance, EuroCRIS[14] aims at "transforming research information into knowledge" while maintaining and publishing the CERIF[15] (Common European Research Information Format) recommendation.

Nevertheless, the major initiatives to concretely build a framework for scholarly communication were launched at national level. The JISC (Joint Information Systems Committee) has funded projects such as Thesis Alive![16] and Daedalus[17] to promote the electronic publishing of theses and dissertations in the UK and the integration of UK institutions in the NDLTD network. SURF (higher education and research partnership organisation for network services and information and communications technology) has supported DARE (Digital Academic Repositories)[18] project to modify the infrastructure of provision of academic information in the Netherlands. However, similar initiatives to create comprehensive frameworks for publishing scholarly material at national level do not exist in all European countries.

The European IST priority on Research Networking (IST 2.5.6) will face the challenge of building a framework for publishing scholarly material, at European level. The DRIVER project (2006-2008) coordinated by the University of Athens will help provide this necessary infrastructure for European research. It will be based on the open infrastructure proposed in the scope of the DELOS network of Excellence for digital libraries[19].

In practice, European actors have extremely diverse administrative organisations, inherited from the past. Interoperability between national systems will have to deal with the heterogeneity of the structures of academic and research entities, their dependencies and relations (as detailed below in section 4.2). Additionally, the implementation of a European framework for e-research will have to face the challenge of multilingualism, with particular impacts on metadata creation and the management of terminologies.

2.3 The francophone context

Francophone e-research networks also face both organisational and linguistic challenges. For the most part, francophone countries (more than 50 countries over 5 continents) are outside Europe. They have extremely different research infrastructures. Several institutions contribute in structuring this community. For instance, directly related to theses and dissertations, the "Organisation Internationale de la Francophonie" (OIF)[20] has funded Cyberthèses (see section 3.1). Several institutions such as the "Agence universitaire de la Francophonie" (AUF)[21]and

[13] <http://www.oaforum.org/>
[14] <http://www.eurocris.org/en/ >
[15] < http://www.cordis.lu/cerif/home.html>
[16] < http://www.thesesalive.ac.uk/ >
[17] < http://www.lib.gla.ac.uk/daedalus/ >
[18] < http://www.darenet.nl/ >
[19] < http://delos-noe.iei.pi.cnr.it/ >
[20] < http://www.francophonie.org/ >
[21] < http://www.auf.org>

programs related to research infrastructures such as "Système d'Information Scientifique et Technique" (SIST)[22] are also helping standardising scholarly publishing in the francophone area.

Many francophone countries actually use multiple languages. They need to implement multilingual systems, with classic constraints in the case of Latin languages and more complex ones in the case of the Arabic language for example. The IMIST (Moroccan Institute for Scientific and Technical information)[23] in Morocco will implement a bilingual union catalogue for theses and dissertations[24].

2.4 The French context

The French administrative organisation is particularly complex because of the multiplicity of complementary administrative frameworks (an example will be given further in section 4.2).
In the last 10 years, no ambitious program has been launched in France to structure scholarly publishing at national level. Public institutions in charge of libraries and scholarly communication such as ABES (Association for Libraries in Higher Education)[25] and INIST (Institute for Scientific and Technical Information)[26] have essentially initiated operational projects such as an integrated publishing chain from articles deposit to the extraction of key indicators for research. Local initiatives are often disconnected from those operations launched at national level.

As a result, the focus of operations launched by French actors tends to be too narrow to enable the implementation of a digital library for e-research, which would federate scholarly communication at national level.

3. Several structuring initiatives

3.1 Cyberthèses

Cyberthèses was born within a francophone program which was also extended to South America. Cyberdocs, its related platform, is an open source software which supports an assembly line starting from document writing to dissemination and archiving.

The main members of Cyberthèses network in the Francophony are the following: "Universidad de Chile" in Santiago[27], "Université de Dakar" (Senegal), "Université d'Antananarivo" (Madagascar) and the National Institute of Agronomy of Algiers [1].

In the Cyberthèses project each university is in charge of the conversion of its theses and dissertations into an archiving format (e.g. TEI-lite in XML). At "Université de Lyon 2", the electronic registration and deposit are now included in the "charte des thèses" which defines the relationship between the student and the institution. The deposit of a complete electronic version of the dissertation is compulsory. The registration is still done by administration, but a workflow software tool was developed which handles the actual deposit and the electronic management of the document and its metadata (DC[28], ETDMS, OAI-PMH).

---

[22] <http://www.sist-sciencesdev.net/ >
[23] <http://www.imist.ma/ >
[24] Beyond the different alphabet between Latin and Arabic languages we must remember that writing directions are opposite. In several metadata elements like dc:description an Arabic sentence could contain an English fragment.
[25] <http://www.abes.fr/ >
[26] < http://www.inist.fr >
[27] < http://www.cybertesis.cl/,universities >
[28] A "TEI.FR" working group has begun to work on TEI-header to Dublin Core adaptation.
< http://listserv.inist.fr/wwsympa.fcgi/info/tei-fr >

3.2 CCSD: open archive with institutional views

CCSD stands for "Centre de la Communication Scientifique Directe" and aims at promoting direct scientific communication between researchers. Very close to ArXiv's philosophy, HAL's[29] software provides an interface for authors to upload into the CCSD database their manuscripts of scholarly articles in all fields. Most of the French research organisations have set up a global agreement for a common cooperation based on HAL which can offer an institutional view for any participant.

A specific service called TEL (thèses-EN-ligne) is dedicated to facilitating the self archiving of thesis manuscripts, which are important documents for direct scientific communication between scientists. TEL can be harvested through the OAI-PHM protocol and two metadata formats are available: unqualified Dublin Core[30], and a specific CCSD one.
A particular feature of this format deals with formal and precise relationships between authors and affiliations which are clearly identified in deposit procedure. This facility allows the institutional views and illustrates the two main goals of CCSD: open archive with a research management orientation.

3.3 STAR: logistic intermediary between local actors and wider actors.

From 2006, the French Ministry of Education, which is responsible for PhD theses infrastructure, will ask ABES, its bibliographic agency, to set up STAR (Signalement des Thèses, Archivage et Recherche), a new service which will operate as a clearing house.

In the input process, STAR will get theses and related metadata from the institutions entitled to guarantee that the given document is true to the original which has been validated by the jury.

In the output process, the digital theses will be delivered to a national digital preservation system which is handled by CINES (Centre Informatique National de l'Enseignement Supérieur)[31]. In addition, metadata will be converted to UNIMARC in order to be sent to Sudoc union catalogue which hosts the theses national bibliography.
Several complementary services (figure 1) will be offered to institutions of PhD defence:

- Sending to CCSD/HAL and other bibliographic databases;
- Full text indexing in SUDOC[32] (Système Universitaire de DOCumentation) academic portal;
- Building a permanent identifier (URI) and resolution for guaranteeing access in any location to a valid copy of the thesis.

Thus, through a unique deposit, a local institution will be able to provide long term preservation and dissemination by many channels, with a high level of traceability in both scientific and administrative aspects.

STAR does not claim to dispense with specific tools or workflows set up by universities. It is true that STAR will offer a web interface to those universities that don't possess any local ETDs management tool. For the others, STAR will ingest locally generated metadata and document files. These metadata will comply with the French exchange format TEF.

---

[29] HAL stands for "hyper article en ligne" <http://hal.ccsd.cnrs.fr/?langue=en>
[30] < http://www.openarchives.org/OAI/2.0/oai_dc.xsd >
[31] < http://www.cines.fr/ >
[32] < http://www.sudoc.abes.fr/ >

TEF (Thèses Electroniques Françaises) is a recommendation provided by an AFNOR[33] working group *(AFNOR CG46/CN357/GE5).* It aims at offering a coherent and flexible organisation for rich and normalised theses metadata: bibliographic metadata (DC), rights metadata (METS Rights), administrative metadata relative to the diploma and preservation metadata. Within TEF, FRBR[34] (Functional Requirements for Bibliographic Records) model is used as a conceptual tool to untangle the notion of theses, METS as an XML wrapping to bind the various metadata modules, Schematron[35] as a precise and flexible validation tool to enforce the business rules that come from the French context

STAR, as a tool, like TEF, as a data structure, plays as a go-between for the benefit of those that produce and authenticate the theses and their metadata as well as for those that make use of them.

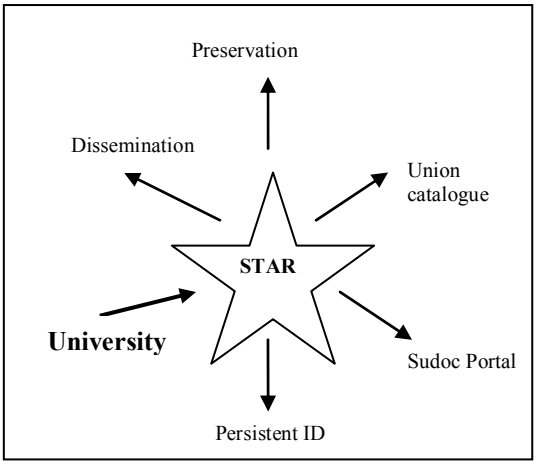

Figure 1

3.4 INIST: Metadata homogenisation

INIST (INstitut de l'Information Scientifique et Technique) is a documentary centre which produces bibliographic databases (Pascal and Francis).

This activity is in permanent evolution. Until fifteen years ago, bibliographic records were manually produced in IS0 2709 format. In a first step, an equivalent SGML DTD was used in order to modernise the production process. Now INIST aims at metadata homogenisation towards a Dublin Core compliant xml schema (Exodic) with automatic indexing.

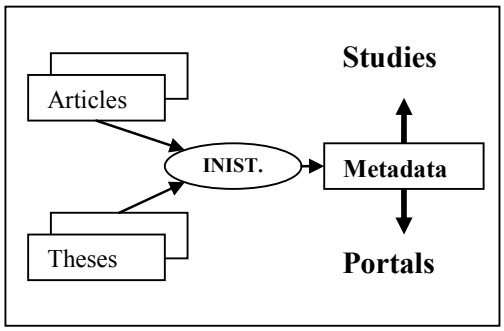

Figure 2

One of INIST's departments is specialised in building thematic portal or handling statistical studies dealing with research policies (Figure 2). This entity is more and more implied in

[33] AFNOR is the French member of CEN and ISO and responsible for all the tasks assigned to France in this respect.
[34] < http://www.ifla.org/VII/s13/frbr/frbr.htm >
[35] Schematron is an XML structure validation language using patterns in trees.
< http://xml.ascc.net/resource/schematron/ >

defining institutional indicators, bringing INIST, like CCSD, to improve the quality of metadata related to relationships between authors and affiliations.

4. Three case-studies

We have just introduced a set of operators which seems to offer a complete set of library oriented services. But, for historical reasons, they had been created quite independently. Thus the reality could be "less than perfect". This paper is written by several people, coming from these organisations, who have realized that interoperability was an important issue, and who are working on exchange formats, generally based on qualified Dublin Core. Is this approach sufficient?

We will now present several case-studies in which we go beyond the basic bibliographic needs (deposit and retrieval) in order to introduce some research policy oriented needs.

In the first case we will analyse the handling of a PhD thesis from the start until its accessibility via OAI-PMH. The "previous designed" workflow shown in Figure 3 looks simple: a thesis is managed by Cyberthèses, and then sent to STAR, and at last to CCSD to be integrated within articles flow. We will consider a situation in which two initial organisations, university and research institution (EPST, Etablissements Publics d'Enseignement et de Recherche) are concerned.

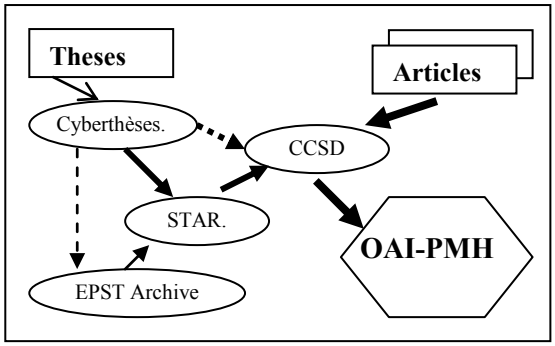

Figure 3

Then we will study two cases of using metadata for doing institutional surveys or creating thematic portals. We will suppose that metadata could have been indexed or upgraded by a documentary centre such as INIST.

4.1 Creating metadata: thinking about reusability

As mentioned before, French research institutions encourage researchers to deposit their work in an open-access repository for greater visibility and added value of their scientific production. These repositories include various types of documents or data: some published articles but also expertises, reports, courses, lecture notes, conference papers, thesis, software documentation or primary data (demographical for instance).

This repository is a significant and valuable source of information for the evaluation of researchers and their research units based on their scientific production. For instance, researchers have to provide information about their scientific production to the evaluation committee every four years. Why do the researchers or their units have to provide this kind of information already available in the repository?

INRA, like any other public research institution, could be willing to transfer data from document repository to the application managing the evaluation files of the researchers and research units. But, even if the metadata provided from the repository is useful, the evaluation process requires new and more complex metadata.

To simplify the study, let's look at the most fundamental exportable data which are required for the "evaluation" application.

> **People:** the researcher who has completed his/her PhD thesis in a research institution has to be identified by the evaluation committee as a researcher, a former PhD student and author of a thesis available in the repository. The researcher may have changed his/her name. Identifying the various statuses or names of a person in order to establish correspondences requires enriching the metadata related to persons.
>
> **Structures**: the research unit where the PhD student worked may be different from the unit he is working for later on as a researcher. Both structures are entitled to claim the search results presented in the PhD thesis, the first one as research work financial support and the second one as researcher's affiliation. The "evaluation" application needs to identify the structure the way it was mentioned in the PhD theses with its equivalent in the institution structural network
>
> **Partners:** the variety of research institutions in France urges to set up a list of scientific partners and to describe the various collaborations. In this way, the PhD student's enrolment university is mentioned in the thesis. The list of partners and/or the collaboration type will have to be completed.

4.2 Institutional surveys

INIST experience shows that detecting relationships between research communities appears to be a key point for research policy [3]. About PhD theses, the computation of affiliations of jury's members could set up several kinds of interesting indicators, for instance dealing with "hidden" research communities.

On a technical point of view, the problem is to extract from metadata several homogeneous items, dealing with people or affiliation. It could be easy in a standardised world; but in reality, a given institution could appear in a quite large number of different lexical forms.
In this purpose, authority files and terminological tables play an important part in the normalisation of the bibliographic data before being handled in the computational process.

In a first step, the authority files can be used to establish the correspondence between well defined items, for example, the names of countries. The technique generally used to establish the equivalent terms and normalise the data fields containing data which differ in terms of typography (upper case or lower case, etc.) or flexion (plural, singular), is to find a convergence to a simpler form, similar to a key with which the given form is associated.

For the majority of indicators, the analytical unit (the object of the study) is a geographical or institutional entity. Publications are assigned to these entities on the basis of an analysis of the addresses of the authors. Variations in the way the names of countries are written are numerically limited. Relating publications to institutions is a much more difficult task which cannot be achieved by a simple analysis of the addresses of the authors appearing in the publications. Very often, a wide variety of different lexical forms for a given entry is found.

This presupposes the existence of geographical (postcodes, towns, regions, countries) and institutional (code for the institution, classification of the organisations by sector, etc.) authority files.

As far as we look for a merely statistical indicator, with a medium quality, this kind of post process is sufficient. But if we need precise computations, some very complex situations could appear.

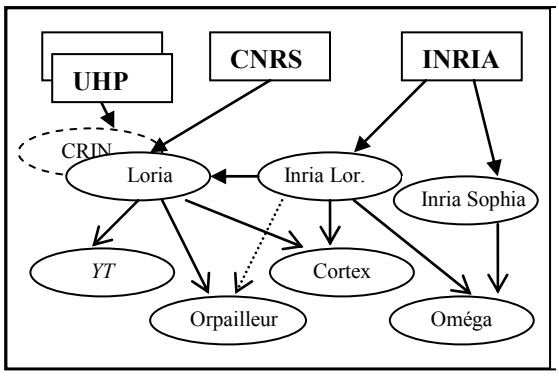

Figure 4

Figure 4 illustrates a realistic situation in Nancy geographical area, as it was two years ago. This example would just illustrate the complexity of what we could meet in practice. A single list of affiliation names is not sufficient and some more sophisticated tools are requested.
At the upper level, we have two government funded research institutions (CNRS and INRIA) and one university (UHP)[36].

At the medium level (research unit), Loria is a joint unit from UHP, Inria, CNRS and two other universities. INRIA Lorraine is the name of the INRIA component in Nancy.
Some years before, CRIN was the acronym for a joint unit between CNRS and Universities but without INRIA. CRIN does not exist at the present time, but a lot of papers or theses are indexed by CRIN and must be handled if we need an "historical study".

At the bottom level (project team), most of the teams, like Cortex, are part of all upper organisations. But things could be more complicated! For instance, *YT* (for Young Team) is only recognised by universities and not by INRIA (and thus by Inria Lorraine). Orpailleur is getting recognised by INRIA. Oméga is a joint team between Inria Loraine and INRIA Sophia but not with Loria.

A consistent metadata schema, such as LEAF[7] one, could offer a solid base which must be completed by a strong study of affiliation links. An ARTIST working group intends to work on this kind of relationship, by using for instance several links which could be issued from something like a "taxonomy of affiliations".

4.3 Thematic survey about biodiversity

This last case study deals with a more thematic aspect of a research policy survey. We have chosen to speak about biodiversity which is becoming a strategic issue. For instance, the European Commission launched BiodiverSA[37] which aims at "setting up efficient trans-national co-operation in the field of biodiversity research funding".

[36] Figure 4 gives a simplified view of the real situation and two others universities (Nancy II and INPL) are concerned.
[37] <http://www.eurobiodiversa.org/ >

In this section we will study a topic which is not really on the agenda of this project but which is considered as very close to its targets: how are distributed the activities of public research laboratories with regard to the main axis of biodiverSA members? This information is supposed to be contained in the publications and more specifically in the PhD theses.

In order to illustrate the complexity of the problem, here are the figures on R&D biodiversity funding in Europe:

more than one hundred funding agencies[2];
several programs by agency, so several hundreds of programs;
several project by program, so several thousands of projects;
several results, such as theses, reports or articles by project, so ten thousand publications!

BiodiverSA intends to create an inventory of all existing biodiversity research funding programs which will be implemented in a "metadatabase" (on a CERIF[38] basis). Vocabulary aspects will play a fundamental role. More specifically classification (or taxonomy) tools must be used with some computational constraints in order to produce a set of indicators.

The BiodivERsA classification *s*cheme is still being designed and it could be composed of three parts.

A general scientific component based on ASRC (Australian Standard Research Classification). ASRC is tightly related to the Organisation for Economic Co-operation and Development (OECD) Proposed Standard Practice for Surveys of Research and Experimental Development.
A specific component dedicated to biodiversity, built with a combination of several existing classifications;
A complementary indexation based on keywords extracted from the "CBD Controlled Vocabulary"[39].

In this context, how could we handle the main topic of this study and, for example, how build an indicator based on PhD theses?

A first problem is to feed a CERIF compliant database which could be used by BiodivERsA with something close to qualified Dublin Core. But the most important issue is the mapping of the resources in the classification system. We could imagine that a few research laboratories will use this classification system in order to be visible by funding agencies. In this case, the indexer needs also to be cautious with the future computational usage of its elements. (This is not the same that archiving or making browsing easier).
But a very large amount of theses related to some particular aspect of biodiversity will not use this schema and several terminological adaptations will be requested. They could be quite easy if the theses are indexed with a well known vocabulary (MeSH for instance).
In the other cases, a document content linguistic analysis should be done. Once again, several vocabulary oriented resources are needed, and this last sample would illustrate the need of complementary terminological repositories.

[38] < http://www.cordis.lu/cerif/ >
[39] CBD stands for "Convention on Biological Diversity".
<http://www.biodiv.org/doc/lists/cbd-voc.pdf>

5. Conclusion

For this new experience (the writing of this paper), after the translation of "*What Is a Digital Library anyway, anymore*", we have chosen to work again from a quite technical point of view. We have identified a large set of stuff[40], such as theses metadata, affiliation links, vocabulary items, which could upgrade our services. We have underscored the fundamental role of a set of repositories of various items and naming conventions which should complete the classical bibliographic archives…

But "*what do we really want to do anyway, anymore*?"

Our common objective is to go further in the e-research or e-science movement and to consider scientific and technical information regardless of the global needs of the research organisations. As we are working in separate institutions which manage different objectives or priorities, this job was not an easy one. Perhaps our most interesting result concerns the identification of all compromises that we have to work with:

1. Compromise between the national environment of theses and the international network.
2. Compromise between the different practices of various actors to ensure reusability of metadata through many applications.
3. Compromise between the needs specific to every kind of users: librarians, informetrics engineers, policy actors, social aspects in networked collaborations (with a particular point about evaluation: indeed the thesis status guarantees a validation process which is the last step of semantic web).
4. Compromise between a focused look on theses and their integration in a larger environment which goes beyond the basic role of a library, even with a "digital" attribute.

In summary, we would consider the theses as nodes within a constellation containing "articles, dissertations, affiliations, vocabularies", but also "patents, projects and numerical results"; in other words all components of a CRIS (Current Research Information System) [8]. Because of a current lack of French or francophone federative research programme, such as NSDL, ARTIST is trying to set up a place where field actors could experiment and exchange information about new practices in producing Scientific or Technical Information.
We would like to consider this paper as a step towards a more regular activity. At present step ARTIST's services look like a "collective scientific blog" and now we intend to produce a francophone electronic journal with peer review mechanism, "electronic style" and sophisticated standardisation. The French language is not to be considered as a "limitation" and we think that new concepts must be grown deeper in a native language training area before international confrontation.

In this context, metadata experimentations give us a natural workshop for collaborative activities that we intend to carry on in the framework of DCMI.

6. Acknowledgements

Only the main contributors are listed as authors of this paper. We would like to acknowledge several other people who have contributed by giving some advice or information (Francis

[40] The translation of stuff, as used in Carl Lagoze's paper, was strongly discussed in a forum:
< http://artist.inist.fr/article.php3?id_article=250 >

ANDRE, Catherine MOREL-PAIR, Clotilde ROUSSEL and Pierrette PAILLASSARD from INIST; Amos DAVID from LORIA, Daniel CHARNAY from CCSD; Ghalia MRAHI from IMIST; Estelle BALIAN from "BiodivErSA – Belgium Biodiversity Platform", or helping in the translation or revising process (Marc RUBIO and Catherine GUNET from INIST).

7. References

1. Y. Bakelli and S. Benrahmoun. Long-term preservation of ETDs in Algeria: discussion through the CERIST Deposit system. In *Proceedings of ETD2003*. Berlin 2003. <http://edoc.hu-berlin.de/conferences/etd2003/bakelli-yahia/HTML/bakelli.html >
2. BiodivERsA. Compendium of Biodiversity Research Funding Agencies in Europe <http://www.eurobiodiversa.org/rich_files/attachments/Compendium%201%20Feb%202006rev.doc >
3. L. Grivel, H. Fagherazzi, P. Fourneret and A. Zerouki. La conception de bases de données infométriques hybrides : analyse de la pratique de trois observatoires européens. In *Journées SFBA proceedings* Ile Rousse 99.
< http://archivesic.ccsd.cnrs.fr/sic_00000464.html>
4. S. Harnad. Publish or Perish - Self-Archive to Flourish : The Green Route to Open Access. In *ERCIM News* January 2006
<http://www.ercim.org/publication/Ercim_News/enw64/harnad.html>
5. D. Le Henaff and C. Thiolon. Gérer et diffuser des thèses électroniques : un choix politique pour un enjeu scientifique. In *Documentaliste - Sciences de l'information*. 42(4-5):272-280. October 2005.
6. Institute of Higher Education . Academic Ranking of World Universities - Shanghai Jiao Tong University, 2005
< http://ed.sjtu.edu.cn/ranking.htm >
7. M. Kaiser. New Ways of Sharing and Using Authority Information. In *D-lib Magazine*, September 2001
<http://www.dlib.org/dlib/november03/lieder/11lieder.html>
8. K. Jeffery. CRIS + open access = the route to research knowledge on the GRID. In *71st IFLA General Conference and Council proceedings*, Oslo, Norway, 2005
<http://www.ifla.org/IV/ifla71/papers/007e-Jeffery.pdf>
9. C. Lagoze, D. Krafft, S. Payette and S. Jesuroga. What Is a Digital Library anyway, anymore? In *D-lib Magazine*. November 2005.
<http://dx.doi.org/10.1045/november2005-lagoze>
10. C. Lynch. Where Do We Go From Here? The Next Decade for Digital Libraries. In *D-lib Magazine*, July 2005
<doi:10.1045/july2005-lynch >
11. M. Patel, Fourth Open Archives Forum Workshop In Practice, Good Practice: The Future of Open Archives, *Ariadne Issue* 37, Oct 2003,
< http://www.ariadne.ac.uk/issue37/oa-forum-ws-rpt/#6>
12. M. Smith, M. Bass, G. McClellan, R. Tansley, M. Barton, M. Branschofsky, D. Stuve, and J. H. Walker, DSpace: An Open Source Dynamic Digital Repository, *D-Lib Magazine*, 9 (1), 2003.
<doi:10.1045/january2003-smith>.
13. H. Suleman, A. Atkins, M. Gonçalves, R. France and E. Fox. Networked Digital Library of Theses and Dissertations, Bridging the Gaps for Global Access - Part 1: Mission and Progress. In *D-lib Magazine*, September 2001
< doi:10.1045/september2001-suleman-pt1>