ACE

Academic Content of Eire

Report about the second year of the IReL project
-abridged version with undisclosed sections-

Leo Waaijers ; Maurice Vanderfeesten

leowaa@xs4all.nl ; vanderfeesten@surf.nl

September 2008

Appendices

1. Portal consultant specification outline final-1 *Confidential*
2. Project plan *Confidential*
3. Consultancy contract *Confidential*
4. a. Proposal CQ2 *Confidential*
   b. Proposal CNR *Confidential*
   c. Proposal Infrae *Confidential*
   d. Proposal Pentila *Confidential*
   e. Proposal PKP *Confidential*
5. Copyright angst, lust for prestige, cost control; Leo Waaijers.
6. Persistent Identifier, Digital Author Identifier; Maurice Vanderfeesten.

7. Usage and citation metrics; Maurice Vanderfeesten.

1. About this report

A. Synopsis

Preamble

The Irish University Association Institutional Repository Project has entered its second year in which a national harvester and a national open access portal for published Irish research have to be realised. In the first year of the project OAI-PMH compliant institutional repositories have been established in four universities. The remaining three universities will follow suit. In the second year the metadata will be validated for compliance with quality guidelines set by the Project Steering Group. Validated metadata will be harvested, aggregated in a national metadata store and exposed in a national portal. The portal functions must be embeddable and will be accommodated within the Expertise Ireland website.

A separate contractor will be engaged to install and maintain the infrastructure which must be based on well documented open source software supported by an active community. The Project Steering Group will decisively oversee the process

The Project Steering Group seeks expert consultancy for the functional specification of the services and the selection of the contractor.

Required services and infrastructure

The national portal service will offer integrated bilingual search and faceted browsing by author, institution, research funder, year of publication, document type and key word (tag cloud). Result sets can be sorted by author (alphabetical order), key word frequency and publication date. The portal will offer an RSS feed service based on search and faceted query. Accessibility (e.g for visually impaired) should

be taken into consideration. A performance of 100 parallel queries per second must be reached.

The following infrastructural services must be realised: harvesting process information and control, validation, normalisation, author identification, collection building, a national OAI-PMH metadata proxy service and website usage statistics. A content management system is desirable.

Consultant's report

The consultant will deliver a report that specifies the required and desirable services and indicates foreseeable post-project developments (i.e. future options like handling enhanced publications, identification of authors and objects, citation and usage metrics, web 2.0 facilities, etc.) The report will outline applicable architecture, implementation requirements, maintenance issues and organisational recommendations. The report will conclude with a shortlist of qualified potential suppliers for the installation and maintenance of a sustainable (and future proof) infrastructure. *---Undisclosed---*

Reference sites

IREL-Open Collaborative workspace of the IReL-Open Project Working Group.

Expertise Ireland The Irish Knowledge Portal.

DRIVER Search portal The European open research publications portal.

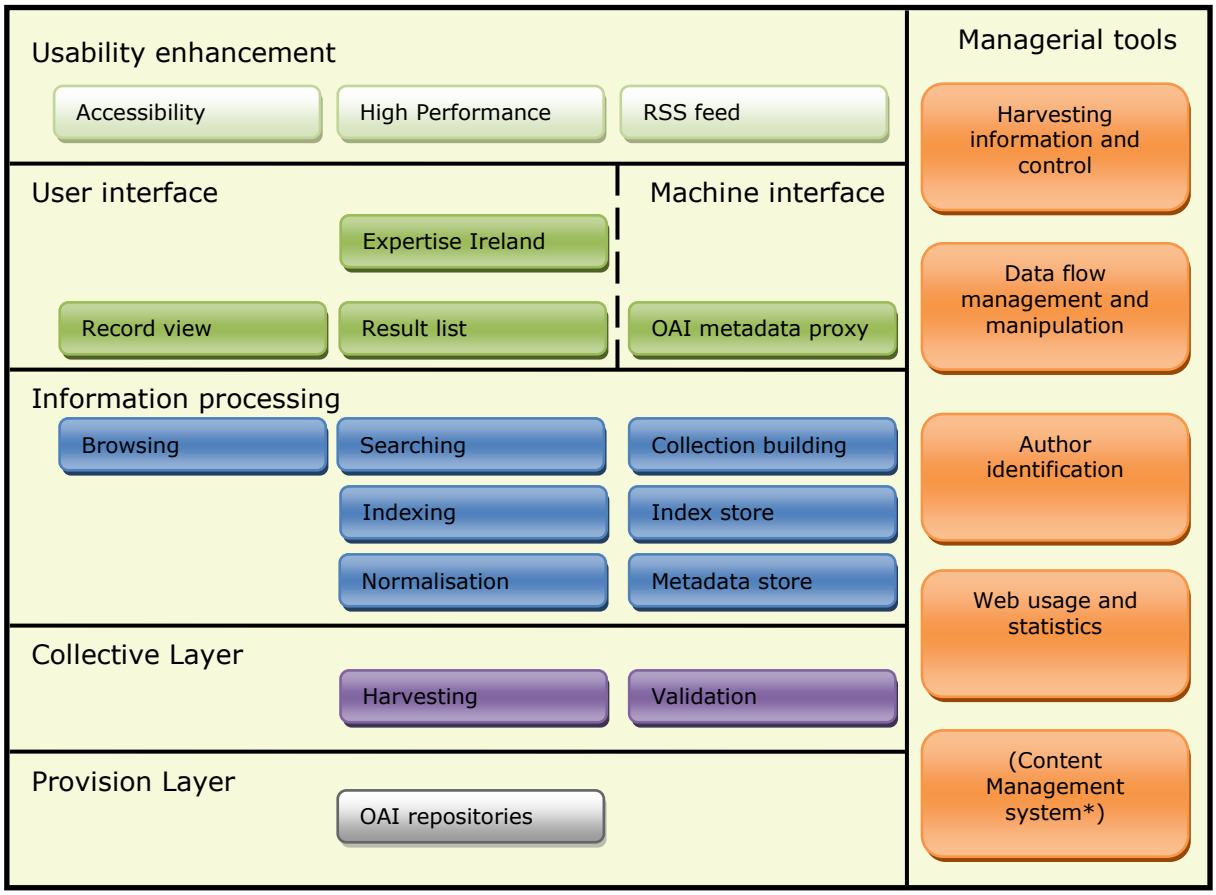
B. Chronology

---Undisclosed---

C. Rights

2. What? IReL deliverables

A. Design



*Desirable service, all other services are required.

B. Functional specifications

NB. The aspects 'content' and 'quality' in this paragraph refer to bibliographic records i.c. metadata fields, not to the content and quality of the underlying publications.

Provision layer

OAI repository	Institutional e-archive containing DC formatted, DIDL packaged bibliographic records of openly accessible completely disembargoed publications. It complies with the OAI-PMH protocol.

Collective layer

Harvester	Mechanism that handles the OAI-PMH requests and responses. Is flexible so can deal with irregularities
Validation	Provides feedback for repository managers to improve their repositories technically and content wise, also flag faulty records. • Reports per repository on general OAI-PMH 2.0 compliance. • Reports per repository on DRIVER specific OAI-PMH rules. • Reports on specific metadata rule failures, per repository pointing the records that fail according to rules set by the Steering Group like DC qualifiers and DIDL packaging. • Flags records with valid metadata for ingestion in the OAI-PMH metadata proxy and for end user quality searches in validated records only. Click here for an example (it is recommended to re-use the DRIVER validator)

Information processing

Normalisation	• Transforms non-DC metadata to the DC format • Transforms faulty content to the correct content

Metadata store	Provides storage of metadata records, original and normalised. • Stores (incrementally) harvested records with provenance information. • Detects and updates for modified and deleted records. • Sends event notification to dependent services (e.g. index engine/store).
Indexing	Metadata from the metadata store with normalised metadata is being processed by an index engine to create an index table. This index table is being stored into an index store. The index service is also able to update changes in the index store when metadata in the metadata store has changed (deleted records etc.) Not only is the metadata indexed, but also provenance information from the OAI-PMH xml. (Enables sorting on repository)
Index store	The index store defines the location of the incoming index information from the index engine. This index store is being used by the search engine for query search operations.
Searching	Uses the index table to perform a search query it gets from the user interface. The result is given to the Result list service. (Standards used: CQL and SRU/SRW)
Browsing	Creates collections on the fly, for example records per institute, subject or document type. Uses the index table to perform a browse query it gets from the user interface. The result is given to the Result list service. (Standards to be used: CQL and SRU/SRW) Enables the user to narrow down or filter the search result to a specific field value for authors, institutions, research funders, year of publication, document types (e.g. e-thesis) and key words. Keywords must be

	presented as a tag cloud. Click here for an example. | Click here for another example
Collection building	Creates default collections. Enables the user to select default collections, such as certain topics like humanities, physics or e-theses. Click here for an example.

User interface

Expertise Ireland	All portal functionalities must be embedded in the Expertise Ireland portal. This is a portal centrally being used by Irish researchers.
Result list	Produces the lists of search and browse results. This can be title, key words, date, author name and institute logo of the records. The Result list allows the user to select the number of records to see (default 10), and to sort the records by date, or author name. In the Result list an end-user can click on the author name, which results in a list of all open access publication that have been produced by this person. (Example) This is based on Digital Author Identification (DAI), where under water the DAI is used to query for records where the author appears.
Record view	When the title of a record in the Result list is clicked, more extended bibliographic information like abstract and usage rights are provided about the record. Required is the redirect function to the publication in the repository, mostly it is the URL in the dc:identifier field. This service can extract the selected data from the metadata store. The record view management interface tells this service where the metadata store is located and tells what data to extract for the presentation to the user.

	Being in the Record view list, clicking on the author name results in a redirect to the Expert profile page (Example) based on the Digital Author Identifier (DAI) We make the assumption that the administrators of the Expertise Ireland portal can make a link to open access publications on the Publications tab on the Expert profile page based on the RSS feed that queries for the author id.

Machine interface

OAI Metadata proxy	Exposes the collected normalised and validated metadata via the OAI-PMH protocol. This service is able to create sets and define the metadataPrefixes based on metadata formats and transformations. • Service that provides locations to store the metadata record, original and normalised. • Store new records • Detect and update modified and deleted records • Event notification for dependent services (e.g. index store) This is a b2b service. Typical users of this service are other service providers like DRIVER, international theses sites and publishers of subject based overlay journals.

Usability enhancing

Accessibility	All Irish websites funded by governmental programs must comply with the Irish web accessibility standards, this means web accessibility for visually impaired persons.
High performance	Since the functionalities are being embedded in the Expertise Ireland portal, they will be used by numerous visitors. The interface must handle 100 parallel queries in one second and come back with a search and browse result

	to the user per query within 2 seconds. The browse function must be fast enough for a user to intuitively "play" with the data he/she is looking for.
RSS feed	RSS feed: The user is able to keep track of the most recent publication he/she searched or browsed for. This is based on the search and faceted browsing query. Access to the RSS feed is placed on the Result list. For an example click here. This service is dependent on the search and browse query that has been inserted by the user. It gets fed by the Result list service and syndicates the content to popular syndication formats such as RSS 0.9, 1.0, 2.0 and ATOM feeds.

Managerial tools

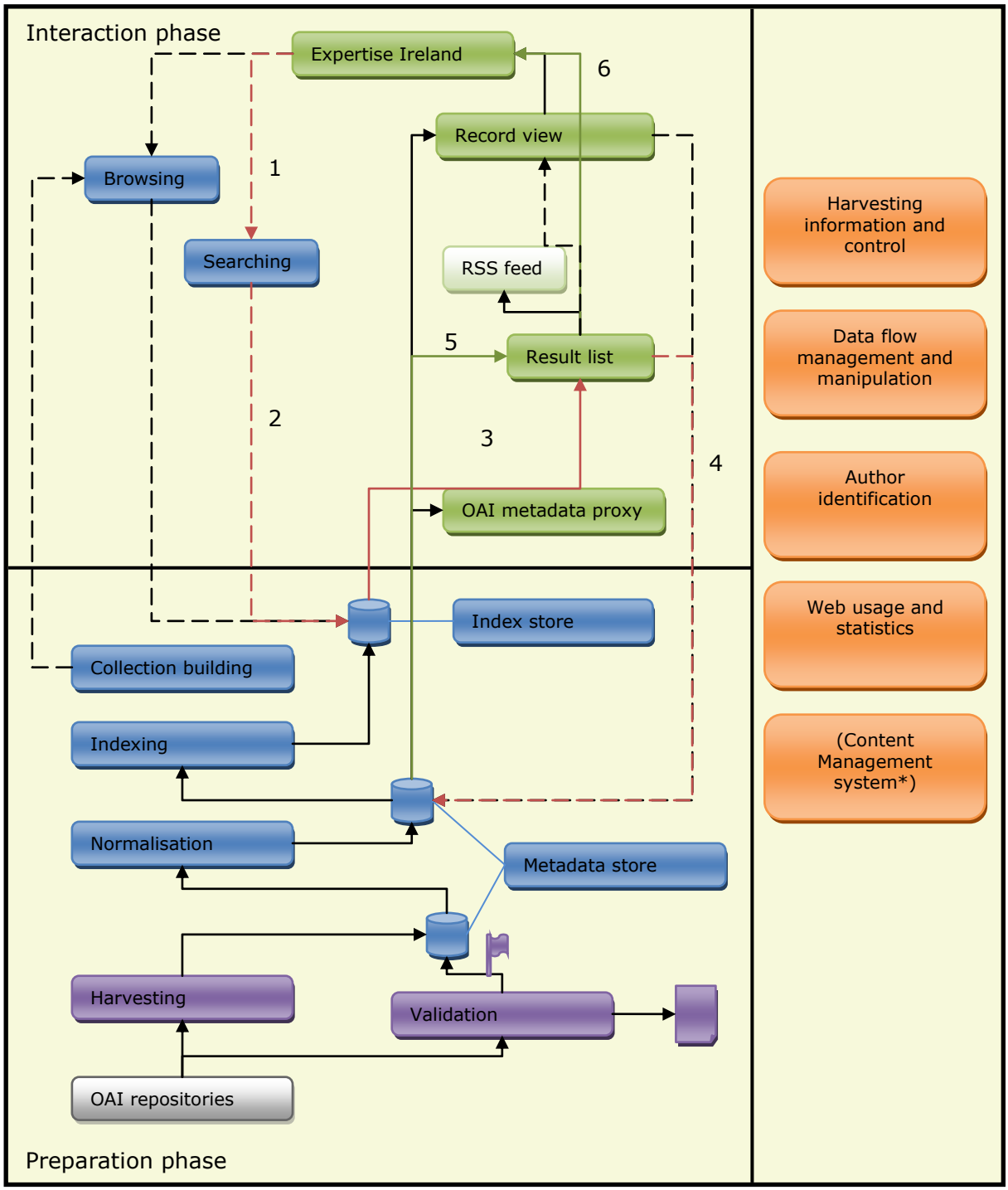
Harvesting information and control	Enables the administrator to add, group and schedule the baseURL's, sets and metadataPrefixes to be used for harvesting. Is web interface to: • Define BaseURL, Sets, Metadata prefix. • Set target to store the incoming data. • Set harvest schedule per repository. Produces information about: • The harvest status per repository. • The harvest result per repository in a graph. • The number of records successfully harvested, harvest errors, repository performance (throughput (records/second), average repository availability and harvest speed). • Validator results (see above)
Data flow management and manipulation	This is the core back-end interface that combines all the management tools that are enumerated below. It enables the administrator to

	• manage the flow of the data in the application • manage the end pages: the controls for the services • arrange front-end presentations for services (like search, browse, collections, rss feeds) **Module tool** to layout the front-end of the portal; i.e. to define what modules appear on what page or event. Such as to let the Browser appear when the Result list appears. Or to show the collections on the front page. **Collection tool** to make collections, based on certain queries, such as e-theses or subjects. **Result list tool** to select and arrange the content that appears on the Result set list. Such as title, institute logo, date, description of max. 50 words, language, author, etc. **Record view tool** to select and arrange the detailed information that is shown when a record is selected by the user. **Browser tool** to select the default metadata fields for filtering. And enables the administrator to select the index table to be used by the Browser service. **Search tool** to select the index table to be used by the Search engine. **Index tool** to select the metadata store to use for indexing, schedule or set an event for indexing and manually start indexing. **OAI-PMH metadata proxy** to define the BaseURL, the sets with metadata records and the metadataPrefix formats. **Metadata store** to store metadata that comes from the harvester and normalisation engine. **Normalisation** to set normalisation rules per repository instance. **Validation tool** to manually validate a repository for

	compliance with specific rules. These rules can be xml, oai-pmh compliance and content rules for metadata. The validator service makes a report for the repository manager, and can be used to set normalisation rules.
Author identification	Within the Expertise Ireland portal every participating researcher has his/her own professional 'home' page. This service makes a binding between the author identifier and the location of the researcher's 'home' page. In this case we make the assumption that the Irish repositories are able to export for each metadata record the author identifier with the author name (when applicable, foreign co-authors might not have an Irish author id) Another assumption is that the researchers in the Expertise Ireland portal also have an author identifier. When clicked on the author name in the Result view, the user is redirected to the Expert profile page where. On this page the user can find the list of open access publications of this author. (Example) Another assumption is that the administrators of the Expertise Ireland portal can make a link to open access publications on the Publications tab on the Expert profile page. This list of open access publications could be generated from the RSS feed from the search application based on the author id.
Web site usage statistics	Interface for website usage statistics. Enables the administrator how many visitors the portal had, what the most popular queries were, what the most popular records are that have been viewed, etc. Important statistics are:

	• Number of visitors per month • Popular user queries • Popular records viewed • Popular full text downloads (link through; links clicked go outside the portal that bring the user to the actual document in a repository) • The number of downloads per document
Content Management System (desirable)	This feature is only necessary when a 'stand-alone' search portal needs to be build where editors dynamically create content. For example news items, blogs, information pages and other CMS features. This could be useful in a testing phase where a front-end framework is needed to work with the preliminary results in a web browser.

C. Interdependencies



All the services work together to present the harvested metadata to the end-user. Therefore they depend on one another. In this section we try to explain the interdependent relations between the services. The solid arrows show the flow of

the (meta)data, the dashed arrows show the direction of instructions one service gives to another.

The illustration is divided in a preparation phase and an interaction phase. In the preparation phase the metadata is made ready for interaction with the user. The interaction phase encompasses the events of the services evoked by user actions, such as a search query.
The red and green arrows depict one case where a user makes a query in Expertise Ireland website. The case will be as follows:

1. The query goes to the search engine (SRU) via the Expertise Ireland portal.
2. The search engine evokes an action on the index table stored in an index store.
3. The records that apply to the query are provided to the Result list.
4. The result list asks the Metadata store to return with a number of selected records.
5. The Metadata store provides the Result list with the data belonging to the records.
6. The Result list provides the Expertise Ireland portal the result (SRW) where the layout and style is applied.

D. Future developments

More web 2.0

Social tagging, rating, bookmarking, annotation	Rate the publication in a 5 star scale (login required, based on Expertise Ireland user management) Annotate the publication (login required, based on Expertise Ireland user management) This feature needs another metadata store to store the social data. The Result list and Record view needs to be changed to combine the data from this store with the metadata from the normalised metadata store. Click here for an example for additional features.

Citing

Record view	Extract and produce metadata cite information in the formats: RIS (compatible with EndNote, Reference Manager, ProCite, RefWorks) and BibTex (compatible with BibDesk, LaTeX)

Enhanced publications

Record view	Enhanced publications are compound digital resources comprising various digital objects like textual documents, research data, visualisations, graphics, models and algorithms. The document that describes the enhanced publication must list all the associated resources, must specify the types and properties of these resources, and must characterise the relationships between these resources. A query produces this document in a human readable (textual or graphical) way.

	Underlying is the emerging OAI-ORE – Object Reuse and Exchange -protocol that describes how to structure an aggregation of web resources by means of a set of pre-defined rdf-triples.

E. Technical annex

An Author Identifier is needed for collecting the oeuvre of an author – that may have been published under different names and distributed over various repositories – and, say, bind this to their expert profile. A Persistent Identifier is needed for the identification of a resource like a publication or a data set. A PI enables, among others, to count the number of citations or downloads of a resource. Pure Dublin Core cannot provide these identifiers in a machine readable way. Therefore the repositories have to make some adjustments in their OAI-PMH output.

MPEG-21 DIDL

DIDL is a structuring format that originally was designed for the audio and video world. Herbert van de Sompel and Patrick Hochstenbach found out that this format was very useful in the Digital Library world as well. Normally we think of an electronic publication in a repository as a solid piece that contains the metadata and the object file. In DIDL this can be considered as a surrogate for the academic piece of work that refers to the pieces of information and data, it is modular (more than one bitstreams and metadata formats) and can grow when time passes (e.g. referring to new converted versions of the PDF format).

DIDL can be implemented by current repository technology and is a pre-ORE solution that does not describe the whole workflow of the academic work, but only the end result, the publication and the parts it consists of.

Below is an example of a DIDL XML file. This can be inserted in the OAI-PMH metadata field. (Click here for the complete and valid example.)

When you look at the example below you will find three <Item> parts inside the DIDL document: the part for the *descriptive metadata*, the *object file* and the *human start page*.

Inside the part for the descriptive metadata the description of the academic work is given in a certain format (DC, MODS, MARC21, etc). This part occurs for every format that has been used.

Inside the part for the object file a link to the actual object is given. These links can be used for future full text indexing, but most importantly it is used by the portal service to redirect users directly to the full text instead of the jump-off page. This

way statistics can be kept about the download usage of the publications. Also this part can occur once or more.

Inside the human start page part the location of the jump-off page is provided. This is for human interpretable reasons where the full semantics are not transported to the service provider, but are only visible in the jump-off page created by the repository software. Also a jump-off page can offer more customised services for the campus user, such as integrated library services.

```
<didl:DIDL>
       <!-- The First Item is the autonomous compound complex entity that is a representation of an academic  work, this also sets the boundaries-->
       <didl:Item>

              <!-- PERSISTENT IDENTIFIER; each academic work has its own global unique identifier, this is being used for resolving and citing to the academic work -->
              <didl:Descriptor>
                     <didl:Statement mimeType="application/xml">
                            <dii:Identifier>urn:NBN:nl:ui:10-6748398729821</dii:Identifier>
                     </didl:Statement>
              </didl:Descriptor>

              <!-- METADATA [1..*]; each metadata format is placed in a separate child-item  -->
              <didl:Item>
                     <didl:Descriptor>
                     <!-- ObjectType of Item, placing Item type descriptors help machines inform what type of information this is, and help separating metadata from a object file and human start page -->
                            <didl:Statement mimeType="application/xml">
                            <dip:ObjectType>info:eu-repo/semantics/descriptiveMetadata</dip:ObjectType>
                            </didl:Statement>
                     </didl:Descriptor>

                     <!-- Actual "By Value" metadata. This is OAI_DC, and could also be MODS, etc presenting information about the academic work -->
                     <didl:Component>
                            <didl:Resource mimeType="application/xml">
                                   <oai_dc:dc>
                                          <dc:title>Neonatal ...</dc:title>
                                          <dc:creator>Bal, M.P.</dc:creator>
                                          <dc:subject>glucocorticoid</dc:subject>
                                                 <!--etc...-->
                                   </oai_dc:dc>
                            </didl:Resource>
                     </didl:Component>
              </didl:Item>

              <!-- OBJECT FILES [1..*];  Each object file (pdf) that is part of the publication is described in a separate child item -->
              <didl:Item>
                     <didl:Descriptor>
                            <didl:Statement mimeType="application/xml">
                                   <dip:ObjectType>info:eu-repo/semantics/objectFile</dip:ObjectType>
                            </didl:Statement>
                     </didl:Descriptor>
```

Mini-MODS

The Netherlands started with the first step in implementing full fine granular metadata format called MODS. The assignment was to transport the Digital Author Identifier along with each record. Therefore inside DIDL there was room for other formats to come along with each record. Mini-MODS was created to transport only the author names, the other metadata elements came from OAI_DC. (Click here for the complete and valid example.)

```
<!—MODS; metadata part to be used in DIDL -->
<didl:Item>
       <!-- ObjectType of Item -->
       <didl:Descriptor>
              <didl:Statement mimeType="application/xml">
                     <dip:ObjectType>info:eu-
repo/semantics/descriptiveMetadata</dip:ObjectType>
              </didl:Statement>
       </didl:Descriptor>
       <didl:Component>
       <didl:Resource mimeType="application/xml">
       <mods>
              <titleInfo xml:lang="en">
                     <title>  Neonatal  Glucocorticoid  Treatment  and  Predisposition  to
Cardiovascular Disease in Rats </title>
              </titleInfo>
              <name type="personal" ID="n1">
                     <namePart type="family"> Bal </namePart>
                     <namePart type="given">M.P.</namePart>
                     <role>
<roleTerm authority="marcrelator" type="code">aut</roleTerm>
                     </role>
              </name>
              <name type="personal" ID="n2">
                     <namePart type="family">Winter, de</namePart>
                     <namePart type="given">R.J.</namePart>
                     <role>
<roleTerm authority="marcrelator" type="code">aut</roleTerm>
                     </role>
              </name>
              <extension>
                     <daiList>
<identifier IDref="n2" authority="info:eu-repo/dai/nl">157455590</identifier>
<identifier IDref="n1" authority="info:eu-repo/dai/nl">123456678</identifier>
                     </daiList>
              </extension>
       </mods>
       </didl:Resource>
       </didl:Component>
</didl:Item>
```

3. Who? IReL contractor

---Undisclosed---

B. Questionnaire

Together with the description of the required deliveries (Chapter 2 of this Report) the following questionnaire was sent to the *---Undisclosed---* candidate suppliers (*---Undisclosed---*). The answers would be taken for their proposals. The next paragraph gives an analysis and conclusion of these proposals.

"The answers to the following questions will be used to reduce the shortlist of potential service suppliers to one or two candidates with whom contract negotiations will be started.

1. Could you please present your organisation in 250 words max.
2. How would you tackle the task?
3. Which of the listed services and functionalities are not included in your current service pack? Would you please indicate case by case which of the following options applies then?
   a. The service will not be offered by you,
   b. You will implement an existing open source solution,
   c. You will develop the service in due course,
   d. You will hire a subcontractor for this service,
   e. Other, namely....

   If c. is applicable, are you prepared to offer the developed software as open source software?
4. For the other services, i.e. those that are in your service pack, would you please indicate per (combination of) service(s) the current state of affairs?
   a. Is the service in use somewhere?
   b. Are you application service provider for this service?
   c. In either case would you mind identifying the user so that we may contact them for references?
5. What organisational/managerial conditions must be fulfilled by the customer for you to deliver your services properly?

E.g. (1) should the Steering Group mandate a representative who can act on their behalf, (2) should the project be subdivided into consecutive stages followed by separate go-no go decisions; in that case which subdivision do you suggest, (3) as only quality controlled i.e. validated records will be processed, should the institutes execute self-validation of their metadata prior to the project, (4) are there specific organisational conditions per (combination of) service(s), (5) other conditions, namely...

6. How much time do you need (for the various stages)? Nb. All the services must be operational in April 2009 at the latest.
7. What technical conditions must be fulfilled by the customer for you to deliver your services properly?

   E.g. (1) what technical competence must be available at customer's side, both at a national and institutional level, (2) should the project have a devoted technical team, (3) what infrastructure in terms of hardware, platforms and applications should be available at customer's side; if appropriate specify per (combination of) service(s), (4) other, namely...
8. About the post-project period: (a) Are you prepared to maintain the infrastructure afterwards? (b) How do you perceive the sustainability of the infrastructure? (c) To what extent will the infrastructure be fit to cope with the future developments that are listed in document 2, section D?
9. What are your financial terms both for the investment/project phase and for the operational/maintenance period thereafter? Educated indications will satisfy for the moment. When it comes to contract negotiations more accurate figures will be required."

C. Evaluation of the proposals

Summary

---Undisclosed---

Technical summary

---Undisclosed---

Analysis

---Undisclosed---

4. How? IReL process

The project is rather complex. Not because the individual tracks are really tough but because it has so many tracks. This project is situated at the crossroads of a growing sense of societal responsibility of universities, changes in information media, an increased interest in the curation of cultural heritage, a move towards sustainable publishing models; and libraries seeking for a viable future. It has the character of a caravanserai where people meet, exchange information, and provision and plan for the coming episode. Contingency management is daily practice here. In such a context it is a bit idle to detail plans. But it is essential to mark critical foresights. This is the intention of this chapter.

A. Timeline

Timeline		Consortial level			Local level		Contractor
1/10/08		Delivery and acceptance consultant's report	Preparing requirements for DAI, MODS, PI and DIDL		Populating repositories & validating against DRIVER validator		
15/10/08		Contract negotiations	Settle IReL Guidelines				Contract negotiations
15/11/08	**M0**	**Kick off event** (see text)	Monitor progress of IReL Guidelines validation and implementation, Monitor population of records		Implementing DAI, MODS, PI, DIDL, IReL Guidelines. Manipulating repository output according to validation reports (IReL GL)		Sign contract, Build initial components
15/1/09	**M1**	**Metadata proxy server** (see text)					Active components: Harvesting, Validation (IReL GL), Normalisation, Metadata Store, Harvesting information and control, Dataflow management and manipulation, OAI metadata proxy.
1/3/09	**M2**	**Demonstrator**	Monitor popularion of records				Active components: Indexing, Index store, Searching, Result list, Record view, Author identification. Extending component: Dataflow management
21/4/09	**M3**	**IReL Open**	Grand opening				Active components: Browsing, Collection building, RSS feed, High performance, Accessibility, Web usage and statistics, Expertise Ireland (embedding). Extending component: Dataflow management

B. Critical decisions

Critical initial Steering Group decisions that have to be taken are:

1. Approval of the consultant's report and start of the negotiations with contractors on the shortlist.
2. Settlement of IReL Guidelines.
   We suggest adopting the DRIVER Guidelines but mandating both the (currently recommended) use of the MODS scheme for metadata and the DIDL container for packaging.
3. Adoption of the DRIVER Validator 'Plus'.
   By 'Plus' we mean that the DRIVER Validator will be enabled to also check for the correct use of MODS and DIDL and for tagging validated records. These facilities have to be supplied by the contractor.
4. Adopt the above timeline (or an amended one) for the second year of the IReL project.

C. Milestones

M 0. Kick Off

The Kick Off should be used as a profiling event. We suggest organizing a working conference with e.g. the following action points:

- Open Access seminar for authors.
- Festive signing of the Berlin Declaration by Irish universities and research funders.
- Signing the contract with the contractor.
- Disseminating an informative leaflet on Open Access and Copyrights.
- Press release.
- Appointment and first meeting of a PR and Awarenss Team.
- Appointment and first meeting, together with contractor, of the Technical Team.
- Start 'autumn school' on IReL Guidelines and Validator.

M 1. Inauguration of the metadata proxy server.

The Irish metadata proxy server will be holding the OAI compliant national collection of quality-tested (i.e. validated and normalised) open access records. The proxy server is the basis for both national (e.g. IReL) and international (e.g. DRIVER, Google Scholar) OA services.

We suggest defining a milestone when every university has contributed at least 1000 publications. This milestone certainly deserves a professional celebration. We think that this should be feasible two months after the Kick Off of the project.

The 'spring school' about the digital author identifier and the persisitent identifier for documents (see below) should be organised somewhere between M1 and M2.

M 2. Inauguration of the demonstrator.

The demonstrator is a spartan web site that proves that indexing, searching and browsing of the content of the proxy server are feasible. The demonstrator must functionally meet all requirements, like faceted searching, high performance and RSS feeds, but these functionalities will not yet be embedded in the Expertise Ireland portal. At a later stage the demonstrator can be developed into a separate portal solely devoted to Irish academic research results coined ACE - Academic Content of Eire – or AWE – Academic Window of Eire - or any other name.

M 3. Embedding in IReL Open.

This is the final stage of the project. In principle this should be technologically easy. Early contacts with the builders of the Expertise Ireland web site must guarantee a smooth integration and prevent technological incongruencies or competence questions.

D. Risks

Repository developments face risks of both a qualitative and a quantitative nature.

"Garbage in, garbage out"

The first risk refers to the quality of metadata and is having both a technological and a bibliographical component. The repository may not be (fully) OAI compliant, the chosen format (DC, MODS, MARCXML, IEEE LOM, etc) may not have been applied properly, the same for packaging or wrapping up techniques like METS, DIDL and ORE, entries in these formats may be empty or not using the applicable (ISO) standards for country, language, time or project based (national) agreements for DAI and PID. A great deal of institutional repositories suffer from these imperfections which are reflected in the precision and recall of end user's search results.

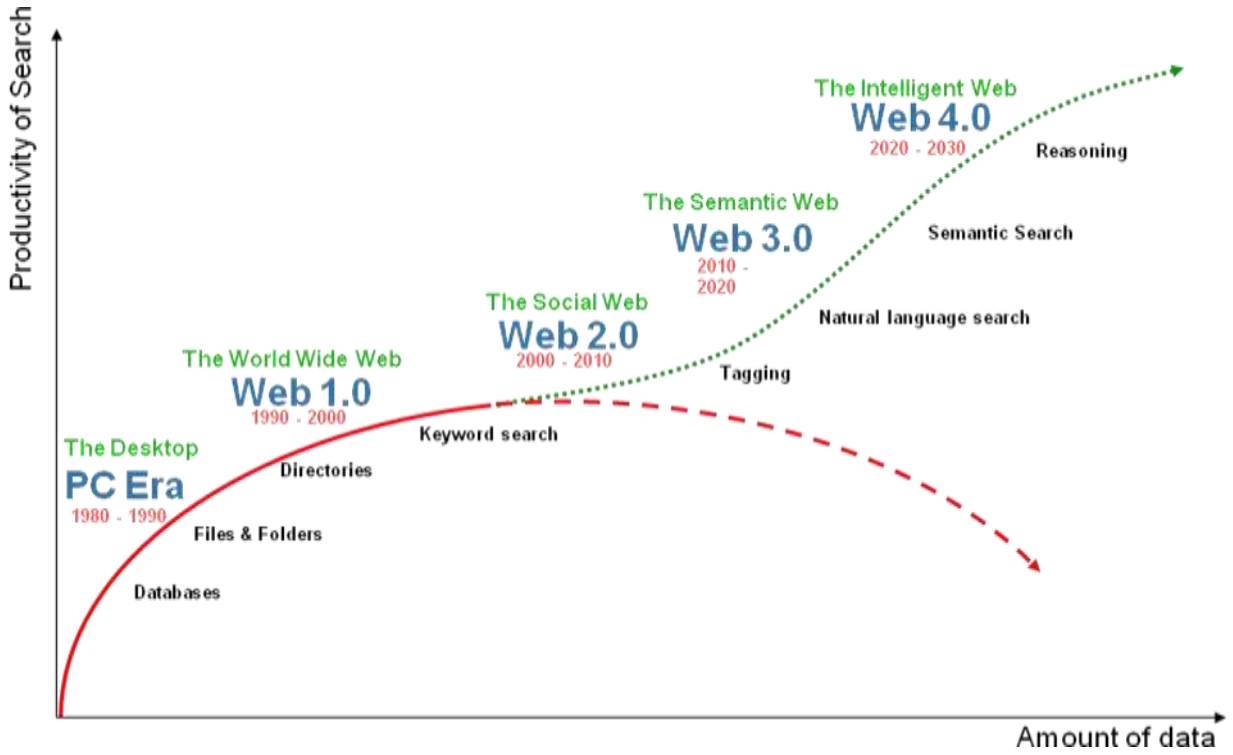
All this can be automatically detected, reported and (partially) corrected via validation and normalisation techniques recently developed in DARE and DRIVER. Adoption of these techniques is included in the IReL project and should be applied as early as possible to avoid huge quantities of metadata that have to be reprocessed.

There are also errors that are less easily detected by computers like typos in names (although digital author identification might help), omissions or poor subject classification. At the end of the day it is classical library professionalism that must guarantee the quality of metadata. Sometimes it is remarkable to encounter classical cataloguing perfection next to modern metadata sloppiness in the same library.

The delivery of the national metadata proxy, halfway through the project, is a main infrastructural milestone. End users may not be aware of this but professionals really have achieved something to celebrate.

In the diagram below we see an illustration that shows the amount of data being produced year after year. The red line indicates the drop in productivity of a search

query when search techniques are not improved and the description of resources is not managed properly. In the coming years, not only are we depending on improved search techniques, but at a basic level we need to ensure that, in our case, the scientific resources are adequately and accurately described. In essence library professionalism ensures correct and valid "tagging" of the resources that fuels the developments towards a semantic web, or as depicted below even the intelligent web.

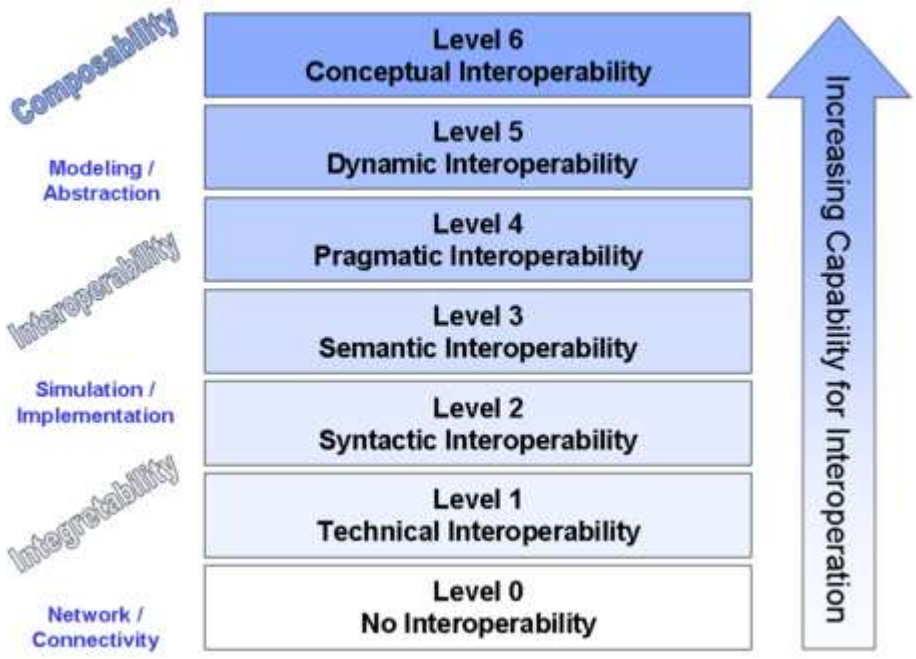


Source: http://www.slideshare.net/novaspivack/nova-spivack-understanding-the-semantic-web-and-twine-talk

In order to reach a semantic or intelligent web that can handle vast amounts of data and make it usable for human beings, one must settle upon basic agreements. These basic agreements are dependent on one another. We need interoperability agreements. The illustration below depicts the levels of interoperability that are needed in order to reach the intelligent web where conceptual reasoning can take place. Currently our repositories reach the level of syntactic interoperability on this scale.

In order to create semantic interoperability one needs to agree on what encodings and vocabularies to use. Therefore the IReL Guidelines are crucial for creating

interoperability on a national scale. In order to be internationally interoperable the IReL Open Project must also adopt international guidelines, like for example the DRIVER guidelines. The DRIVER guidelines may not be perfect, but they are the closest thing the international community has that provides syntactic and semantic interoperability for repositories. To make the comparison to two persons talking: In order to communicate the speaker must somehow be sure that the listener understands what he/she is saying. They need a common ground, agreements of concepts they both understand and of which they are aware.



Source: http://en.wikipedia.org/wiki/Conceptual_interoperability

As a common problem faced by repository staff is being unsure of what exactly should be inputted in each metadata field we suggest that this fall an 'autumn school' be organised for this staff to make them aware of the essential role of high quality metadata and to introduce the IRel Guidelines and the Validator. Later, e.g. next spring, this could be sequeled by a 'spring school' introducing the digital author identifier and the persistent identifier for documents and other more sophisticated developments.

"Nothing in, nothing out"

Even when institutional repositories function as they should – easy to use, well embedded in the local and national infrastructure, and recognized by the local authorities – authors and their managers may still be hesitant or reluctant to use them for offering access to their publications. This then has little to do with technology and metadata but everything to do with culture and costs. These aspects must also be addressed in the IReL Open project on penalty of ultimately delivering perfect but empty repositories. Parallel to the technology task, an awareness-raising programme should be implemented. The festive Kick Off meeting of the project could be the start. Universities should make their support for OA manifest by signing the Berlin Declaration at that occasion. This is the starting point of local OA campaigns underscored by a widely disseminated informative and easy-to-read brochure detailing the legal (copyright), cultural (prestige) and managerial (costs) aspects of OA. Appendix 5 of this report is a memorandum that can be used as a basis for such a brochure. Subsequently, during the IRel Open project a number of prestigious articles of Irish scientists and scholars might be collected in the institutional repositories, harvested and published to prove that prestige and open access is a perfect match. Finally, we suggest organizing an international OA high profile conference at the occasion of the completion of the project, i.e. the accomplishment of an infrastructure that guarantees the future flow of Irish academic research results to the Expertise Ireland portal; 10.000 (or more) openly accessible research publications could be a good starting point.