Institutional Repositories, a new platform in Higher Education and Research

A discussion paper commissioned by SURF

Joost Kircz

Final version
7 July 2005

KRA publishing research

The right question to your solution

Institutional Repositories, a new platform in Higher Education and Research

This discussion paper is commissioned by SURF, the Dutch higher education and research partnership organisation for network services and information and communications technology. It does not (necessarily) reflect the opinion of SURF regarding this issue. The sole aim of the author is to sharpen the arguments.

The discussion on Institutional Repositories is wide ranging. It deals with all aspects of higher education and scientific research. The electronic era reshapes the whole academic landscape, research methods, information management and education alike. In such a dynamic situation, it is important to keep an analytic eye on all processes and stakeholders. We have to be sure which battles are fought, which strategies are implemented and which arms are used. We have to avoid the standard mistake to fight the present war with the strategies of the former or in a less belligerent language to repeat the example of building clippers after its finally successful rival, the steamship, was engaging in transoceanic voyages.

This short report aims to provide an analytical approach toward the issue of Institutional Repositories in order to clarify the multifaceted discussion and to serve as a kind of checklist in the process of decision making.
The present discussion on Institutional Repositories is characterised by an unbalanced mixture of technology push, a discussion on the level of academic autarky and the debate on the role of (commercial) publishers. In this report, a more axiomatic approach is adopted in the hope that this will assist reaching a more transparent discussion of the aims and goals as well as the organisational and financial consequences.

The report has two parts. Part 1 is a long analysis to delineate the various components in the discussion. Part 2 deals with the current debate.

Helpful discussions and critical comments by Kurt de Belder, Liz Lyon, Susanna Mornati, Frank Scholtze, Lilian van der Vaart, Solke Veling, Leo Waaijers and Gerard van Westrienen, are kindly acknowledged.

A report Commissioned by SURF-DARE
Author: Dr. Joost Kircz
KRA- Publishing research
The right question to your solution
www.kra.nl
Amsterdam, 7 July 2005

Institutional Repositories, a new platform in Higher Education and Research

Part 1
Fundamental issues to be defined before we can go on

Part 2
Do it yourself or delegate
The Institutional Repository as a new research platform

Part 1
Fundamental issues to be defined before we can go on

I. The tasks of the Institutes of Higher Education

The starting point of all discussions is the mission of the universities and other institutions of higher education and/or scientific research. In short: scientific research and education, as well as all kinds of communication about these activities within the institutions and towards society at large.
In the context of the present discussion on Institutional Repositories, this boils down to the following **tasks**:

I-1. The **Collecting** of all data, texts, objects, etc., that are relevant to fulfil the mission. This collection comprises on the one hand originals of, copies of, or access to representations of past performance, on the other hand the day-to-day-collecting of current own production of scientific and educational creativity.

I-2. In order to make such a collection useful, it needs **Indexing**. The level and depth of the indexing are part of the current discussion. It goes without saying that at least the present standard metadata fields are a minimum. An administrative tool like METIS in The Netherlands is an example of an attempt to identify written products of scientific research. On the level of content description we deal with the important issue of context dependent controlled keyword lists, thesauri, taxonomies or ontologies, needed to explicate the knowledge and information hidden away in the collection.

I-3. Collecting and naming the total scientific and educational output of an institution is one thing, converting it to a useful source of new activities demands **Enriching** of the collection by permanent adding new metadata in the form of the latest index terms in the particular field as well as dynamically linking the various items in such a way that powerful analogs and metaphors become visible for the prepared mind of the scholar.

The tasks in I-2 and I-3 are closely related to the research in co-called Semantic Webs.

I-4. To enable usage means a professional level of **Storage** that guarantees undamaged retrieval. With storage we refer to the library function, namely that part of the collective memory that is immediately approachable by the working scientist, teacher and student. This is also known as dynamical archive opposite to I-5.

I-5. As academic knowledge is a friend forever, in the sense that we never know when we long for or need a particular item, the **Archiving** of the complete collection is a separate item in the electronic era. The issue of digital preservation is certainly a key concern to guarantee the continuing accretion of human knowledge and data. Here we talk about the static archive.

I-6. All of the above demands a control system that **Guarantees Integrity** of the collection on various levels. In the first place the stored items must be kept as unaltered as possible in order to be inspected, reinterpreted or reused as they are. Secondly no clearing out of material is allowed as a rule (as in administrative systems where files are pruned and kept only as long as the archival laws demand). Versions and mistakes, failed experiments and wrong analyses are part of the scientific discourse and very often old failures turn out to reveal serious insights

that cannot be overlooked. The collected papers of Newton, including his astrological excursions are still available. In an electronic environment, discarding work that is unfashionable work or deemed prejudicial, is much easier.

I-7. Integrity is closely related to the level of certification and validation. **Certification** and **Validation,** are essential tasks of the academic responsibilities. This issue is not only important in the case of scientific journal articles, where an external peer-review system organised by publishing houses, is common practice. Also pre-prints, reports, drafts, data etc. (see below section II) need validation and certification. Proper policies have to be implemented on the IR level.

I-8. Finally, a main task is the **Communication** of what is available and how all this content is approachable and usable for scholars and normal citizens alike. The various concrete activities that fit the vague notion communications have to be made explicit in relation to its stakeholders. On the institutional level it is often the show case and the wide dissemination of what has been performed. In particular the traditional role of the library of providing information to the researcher will be augmented with the reverse role of presenting the local author and her/his works to the world. On the scientists' level it is mainly a matter of finding and to be found, as well as discussion and collaboration. Finally on the teaching level it can be a more fluid interactive dialogue. In all cases the level of completeness and the ways of certification and validation will be different (see section III)

II. The Institutional Repository

Given the tasks mentioned in section I, we now enter the discussion which items belong to an Institutional Repository. The list below is certainly not universally accepted and subject to debate. In different disciplines, different varieties of knowledge representations are used, the role of which can also differ considerably between fields.
In order to assist in reaching an Institutional Repository policy, the following is just a tentative checklist of **kinds** of knowledge representations.

Before we list the kinds of information an Institutional Repository might take up, it is an essential consideration that an electronic Institutional Repository is not only a collection of digitised textual works. An Institutional Repository is an electronic warehouse, filled with digital files. One way of presenting content, stored in this warehouse, might be print on paper, but there is no reason to demand that this kind of representation is compulsory for all items. Some digitally born items refuse to be printed on paper. Institutional Repositories are arsenals of recorded knowledge: not only textual but also audiovisuals, data sets and software.
At this place it is important to note that, although an IR will in principle only host complete works (full text in the case of text), in many cases it will be a mixture of this and pointers to other repositories. This can be because the other repositories have a discipline dependent structure such as in genetics or astronomy, or due to rights or confidentiality issues such as patient records.

II-1. **Educational material.** This broad type of information is ill defined. Lecture notes are obvious examples. Also examination questions (and their answers) and exercises are typical items that are normally reused. But short presentations, (ex)sample collections and other more

transient items have no clear status yet. On the other hand, digitally born learning environments, made for multiple use, are examples of educational material that scream for a clear policy. In particular here we see a conflation of reporting of research findings and presentation. The difference between teaching and learning material is also the stability over time. Distant learning courses must have a certain degree of stability in time as a variety of students are working on different times and with different paces. On the other hand teaching material can easily change during a course if deemed appropriate for the local circumstances. Also the rights question can be different as the goal of teachers and practitioners is different from researchers whose goal is recognition by intellectual ownership.

II-2. **Master and Doctorate Theses** are the most visible products of students. Dissertations are proof of an accomplishment and are the witnesses of past performance. Typically, the usage of most of this type of reporting is delegated to excerpts published as journal articles, thus losing context and background information (e.g., in an experimental dissertation normally a fair share is devoted to explicate the intricacies of the measurements, a description that is normally shorted in a journal publication). Nevertheless, this type of reporting science is very well certified as the name(s) of the supervisors are known (a good example of open peer review) and so is the appreciation in the form of marks or grades. It is therefore obvious that all such reporting is part of an Institutional Repository.

II-3. **Reports and working papers** tend to be so-called "grey-literature", a strange name for often crucial information. Reports can be extensive studies, designs of (large) instrumentation or overviews of any kind. Such reports are part and parcel of the output of all institutes of higher education and research. Although it should go without saying, reports are not yet an integrated part of the institutional public reporting.

II-4. **Annotations** are in many fields, such as law studies, an essential enrichment of the body of knowledge. Their form and structure, and embedding in a larger corpus, demand special attention.

II-5. In medicine and pharmacology **Protocols** become more and more important. But also in social studies protocols are important to enable comparisons and integrity checks. Hence, they also belong to the realm of an Institutional Repository.

II-6. More and more **Data sets** are now part of the reporting of a scientific investigation. In mega (giga,tera) form they exist already in external huge databases such as in astronomy, genetics and other biomedical fields. In the Social Sciences, data sets are essential and unfortunately often very badly stored. The issues here are:
a) To what extent does an Institutional Repository contain data sets. Does it include such academic products or does it only provide pointers to stored data elsewhere? This is not an easy issue as the intellectual property aspects, in particular in the biomedical field, are under permanent surveillance.
b) Many data sets are only useful in combination with others and only the aggregated form is of value. It is tempting to point to the EU database directive in this respect, because here new IPR are granted if value is added on a collection of otherwise rights free information.
c) Although not yet universally accepted, it should be normal that all data that support a scientific claim are permanently available for reuse, and reanalysis in view of the continuing progress in the field. Already now, many journals accept data sets as appendices to a publication.

II-7. **Papers and Preprints** are the most hotly debated information objects in the digital library and Institutional Repository discussion. In fact they are only the reporting of authors, as far as they consider it important to share their insights with others. In the social context of *Publish or Perish* unfinished or abandoned research endeavours are normally tucked away, therewith disabling others to learn from mistakes or too ambitious set-ups. A serious discussion is needed to what extent an Institutional Repository is also the source for trial and error reports (which can be certified as well) and not only a trophy cabinet, as publication lists sometimes suggest to be. As the issue of indexing, certification and the publishing role (not to be confused with the publisher as company) is dealt within section VI, it is sufficient here to pose the question if the Institutional Repository contains a reference or the complete document.

II-8. **Proceedings and conference contributions** are closely related to journal papers but are certainly not the same. Very often they are interim reports of current research and more and more these reports are now only published on-line on a web site.

II-9. **Books** are certainly important candidates for Institutional Repositories. In a draft version they can take the form of worked out lecture notes. In particular in the humanities many books are of interest for the general public and published by (academic) publishers. In such a case it would be best if the source text resides in the Repository, but remains dormant, in other words inaccessible, until the book is out of print.

Concluding we can say that we see a whole gamut of differently styles IRs. Not only on the geographical level (institute, university, state), but also according to discipline (chemistry, mediaeval music, etc.). Another axis will be the coordination of the material, as text, images, and data demand different metadata and sometimes database structures. The discussion on what will define an information object worth its own Uniform Resource Locator (URI) now has the floor. Based on that, the issue of federated IRs and the necessary protocols and standards can be dealt with. In particular we encounter the difficult issue of how to point to experimentally generated data that are not (yet) published, that is to say referred to in the standard literature. As an example it has been reported that related to the UK Crystal Structure Database it is estimated that only 20% or less of the data generated globally is reaching the public domain. This problem has two aspects: firstly, immense amounts of data are siting idly in databases waiting for treatment (such as the results of space craft measurements) and secondly, many research data are simply not properly stored by researchers and lost as soon as a research programme changes direction or even as a person retires or changes his/her jobs.

III. The usage of an Institutional Repository

As was clearly stated in section I, the role of the institutions is to foster and advance research and education in the sciences and humanities. The Institutional Repository is a prime instrument and therefore it is worthwhile to review the different kinds of usage. An important issue is the relationship between the technological data level and the sociological service level. The first level is infrastructural and in a certain sense supply driven, while the second layer is driven by fashion, demand and opportunity. Below we list usage, independent of the technological foundation (see section V).

III-1. An essential task of the Institutional Repository is to store and retain the proof of performed scientific and educational labour: the intellectual sweat of the academic pursuit. This usage is the basis for the following types and in itself it **Accounts** to society what has been done. The digital shop-window of the institution adds to the institutional profile.

III-2. The Institutional Repository is a **Source** for the definition of new works and research projects. It enables a check on what already has been done and on the basis of which we can build further. It tells us what lines of research are popular, over- or underpopulated, and how it relates to other people's works.

III-3. The Institutional Repository enables a check **against Duplication**. An electronic repository enables a far better check on past activities than a traditional paper-based literature search. As mentioned under I.6, a mature Institutional Repository, in contradiction to journal publications, can also harbour (yet) unfinished works. Based on official journal publications it is far more difficult to assess if some ideas have already been tried and failed.

III-4. In close relationship with the previous issue the IR has the task to explicitly present the intellectual ownership, which entails more than the Dublin Core metadata field <creator>. Therefore, the IR plays a role in the **career evaluation** of a person as well as tool for evaluating **research grant applications**.

III-5. The Institutional Repository is a direct **research Tool**, like a measuring instrument. For the humanities, annotated corpora are the laboratory for search, analyses, discovery, and new theory. The same is true for the data sets in, e.g., the social sciences. The plans for data warehousing and the creation of large annotated and annotatable corpora are closely intertwined with the Institutional Repository idea. It is a technical decision to what extent such databases are fully integrated.

III-6. An Institutional Repository is a source for educational **Training**. Not only in the sense of reusing stored courses or lecture notes, but a medium in which research progress, methods and discourse can be analysed to learn from.

III-7. Finally the Institutional Repository is a source for **Reuse** of information. The Institutional Repository allows for integrating bit and pieces into new knowledge and data into new products. In particular if raw data are stored comparative studies, even between projects that are years apart, become possible. On a more trivial level just excellent examples from lecture notes can be called on and reproduced when needed. Reuse and duplication are the two sides of the same coin.

III-8. On top of these more trans-historical usage that is not necessarily fully dependent of electronic media, we have completely **Novel** usage such as the automatic generation of citation scores, the creation of dedicated or personalised websites, and most importantly the creation of subject, quality and level dependent portals, that mimic traditional journals (hence sometimes called overlay journals).

IV. The users of the Institutional Repository

Repositories are built for people, therefore it is important to stipulate what kind of services an Institutional Repository will be helpful for.
We split the users in two categories: the readers IV-1-> 4 and the authors IV-5 -> 8

The readers

Below four roles of readership are defined, in which readership is understood in the broad sense of consuming information and knowledge. It is important to note that every scholar on different occasions will play all these four roles of readership depending on the stage of the research or teaching programme. An important variable is the discipline, as reading in applied mechanics is another style of reading than in linguistics.

IV-1. **The Non-reader**. Although normally not named as a separate entity, the sector of interested non-readers is an important group, consisting mainly of administrators only concerned by the fact that a particular person or research group has produced, where, how often, and what type of product (for example review, book, research note, lecture note).
For this group of users of scientific and educational information, the usual bibliographic information is normally sufficient. The Institutional Repository content serves a role in essentially bookkeeping operations and in most cases does not require reading at all. In the concrete Dutch university environment we are talking about the METIS administrative environment where all publications are reported. A direct coupling between such systems and the Institutional Repository is obvious, but as argued above, the Institutional Repository houses more than only validated and certified publications. For these users metadata like citation scores, number of downloads, etc. are often considered as being important.
In its capacity as research tool, see III-5, the structural data about an IR and the various metadata schema's are research objects for information and computer scientists, who develop descriptive languages, databases, search techniques or deal with long term digital preservation. Also, for these readers, the precise intellectual content is not important.

IV-2. **The Informed reader**. This type of reader is a scientist in the same field as the author, or has at least a good understanding of the field. This type of reader knows what he or she is looking for and is able to find his or her way in the literature quickly. The authors' names are known. Keywords, which are not always very precise, are complemented with background knowledge. This type of reader selects a limited number of items and scans the contents. The queries are clear and the use of standard bibliographic search techniques and keyword systems is sufficient most of the time. The narrow selection of items retrieved by the informed reader just serves the direct, known, information need. Only the hard core of information is used and there are hardly any reasons to read, e.g., the article in its entirety. Reuse (III.7) is a typical goal for this reader.

IV-3. **The Partially informed reader** is someone who is starting up in a field, or is just generally interested in adjacent fields to his or her own. Names are sometimes known from review articles or references, but their importance is not fully appreciated. Selections on keywords normally give rise to too many hits in searches. It is very difficult to find the needle in the haystack. The only hope is to find a good recent review article and to trace back the references. Using a citation index one is also able to search forward by retrieving those articles which refer to a particular 'old article'. For these users it is important to be certain that some work is already done and some avenues are already tested.

IV-4. In the last category we find the **Uninformed reader.** A most interesting clientele for an Institutional Repository. They are the people who want to learn something new. Typical questions might be: 'What are they doing in high-temperature superconductivity?'or 'What type of methods exist in data reduction?'and so on. The quest is not precise.

For the categories IV-3 and IV-4 search engines that include some kind of taxonomies are very important.

A final remark is that the reader can be an academic or governmental scholar or somebody who works for a commercial company. This difference can play a role in the final ruling about free access or access for a fee.

The authors

Authors have to understand that they fulfil the various readers' roles in the process of writing. The usage of the Repository as explained in section III, is essential for their production. Hence, authors belong to the most important group of users.

IV-5. The **Educator and teacher** compose the group that will fill and use the Institutional Repository for sustained re-use and training. In particular for learning on distance programmes their digitally born material represents a new branch of authoring.

IV-6. The **Students**' products in the form of theses of all levels are usually a bit under represented. Some universities now demand that all doctorate theses are digitally available in a repository. This is a good first step, but in the bachelor-master system, in fact all theses must find their natural place in the Institutional Repository.

IV-7. At present the production of the **Academic Bureaucracy** is often overlooked as one of the prominent (though not always by everybody heralded) products of the Institutes for higher Education. For a good account of the institutional's health, these products that determine many essential aspects of the academic pursuit, belong to the Repository.

IV-8. The most important category of authors is the **Researcher** who produces various kinds of material as listed in section II. Immediately it becomes clear that we cannot talk about one type of author. Authorship is most dependent on discipline. In mathematical and particle physics we know a unique culture of preprints, a system that was already in place long before the electronic revolution. In clinical medicine, preprints are often not appreciated as medical information must be certified before the general public gets hold of it. In civil engineering, the publication habits are different again. Most of the time it is in the vernacular and bridges and buildings are more robust proofs of performance than articles. In many fields the experimental results are stored in large international database systems. Published articles refer to such systems. Molecular biology is a good example. They work with biological entities such as cell lines, proteins, genes, species all stored in dedicated databases. They are used to accessing virtual representations of these objects online through sites like Genbank, etc. Here we immediately see the merging of the research and the publication platform. The usage of electronic media changes the way research is executed and reported. The repository becomes a pivot (see also III-5).

V. Technical issues and constraints

Though not the subject of this report, we cannot go further without at least identifying which technical constraints and capabilities constitute the framework in which a discussion on Institutional Repositories is fruitful. Here we deal with the data level.
Section II above provides an incomplete list of possible kinds of items to be stored in the Institutional Repository, defined in terms of products. The items in the Institutional Repository can be very different in character, which demands a clear policy. This aspect induces the issues of federation of IRs with homo- or heterogeneous content and shared resources.

V-1. On the basic data level, an Institutional Repository must adhere to pertinent international **Standards** that allow interoperability, digital preservation and harvesting. In this report no argument is needed to state that open source software and open standards have the preference. This means that all raw products from the institute's staff and students need to be standardised before they are entered into an Institutional Repository. This implies that for all materials a conversion to a (preferably open) standard is needed before it can be accepted for the IR. After all it is of little value if texts appear as images or in an idiosyncratic wordprocessor format. The same holds for the many equipment-dependent image file types in microscopy. Based on a standardisation of file structure and metadata, federated IRs become possible.

V-2. We have **Text** that is to say a string of symbols representing an alphabetic ordering of natural language. These bit stings can be manipulated in word processors and by various kinds of search algorithms. Though old-fashioned ASCII is often still the rule and still useful on a low level of computer infrastructure, it is obvious that if an Institutional Repository wants to deal with more than American English texts, Unicode must be the standard.
Furthermore, text is delivered presently in a great variety of proprietary text processor encoding schemes. The choice has to be made if we want the Institutional Repository to be just a shopping bag of random items, or a storeroom where we can search, manipulate and compile new items out of parts of other ones. In the last case, the text files have to be stripped from their useless embellishments and stored in a clear standard. An agreement has to be reached on the, per discipline relevant, text encoding that can be represented in the present XML standard. The XML approach is a good example for data representation and already an industry standard for the big international publishing houses and companies that produce massive manuals for their industrial products. It would be a strange fact if the Institutional Repository turns out to be just an electronic storehouse of incompatible files. Some of the main roles of the Institutional Repository, namely the preservation and integrity would be violated by every mayor change in word processor fashion (see III.1 & 7).

V-3. **Images** appear in a great variety of encoding standards, many of them not in the open source domain (such as JPEG). Images become more and more important in scholarly work as their sole role as illustration to the text (in the print on paper world) is changing to objects of primary information. In particular in clinical medicine and art history, images are the object of knowledge and the text is the explication of the image. Image handling is therefore an important field in itself. The Institutional Repository itself could be an important corpus for image processing and pattern recognition research. This supports consideration III-5, that the Institutional Repository is a research instrument itself.

V-4. As mentioned already in II-6, **Data sets** belong to the realm of Institutional Repositories. Also here, the definition of formats and database structures is in full development. As more and more academic output will be in the form of data sets, the Institutional Repository is a research object in itself. The important short term decision is to what extent the Institutional Repository will host data sets itself or only contains descriptions of and links to collections elsewhere.

V-5. Just for completeness sake it is good to mention **Streaming media**, as audio and film are now part and parcel of many academic fields, while games are slowly entering the stage.

V-6. The last kind to be mentioned are **Enveloped** items such as PDF files, a continuing changing representation standard that brings together text, images, audio files, etc. in a format that mimics a printed page. Although PDF can be considered as a defacto standard, the fact that the final product is a "Gesamt Kunstwerk" almost prohibits the reuse of it components, a prime demand for the usability of the Institutional Repository. It goes without saying that a PDF file as a representation of a particular product of an author is quite acceptable. But like a well-printed page, it does not allow easy reuse in new works.

V-7. **Educational material**, in particular in digitally born form they are often of a special kind. Games software is now frequently used for instructions and learning environments. The problem of versioning is most relevant in the field of learning and teaching material. For example a blatantly racist course of the 50th of the last century might be replaced since long, but remains a very valuable source of information on how people thought and got educated in that period.

V-8. **Metadata** as such comprises a fair share of the total file. Not only do they describe the content and structure of the material but are also used for digital right management systems that overview access and usage. Metadata are essential for search and discovery software as well as for retrieval purposes.

V-9. **Digital signatures and time stamps** are necessary tools to certify and guarantee the integrity of the work. The caveat here is that an electronic work is meant for reuse and multiple use. As the most important advantage of an electronic research environment is its capacity to merge activities that are disjoint in time and place, we have to return to discussion on the granularity of an information object and the related identification of objects in terms of Unique Resource Identifiers.

V-10. A clear understanding must develop between the protected **archival** sever and the maintenance and services of the information object servers that interact with the primary users.

VI. The Institutional Repository and the publishing en library functions

For a transparent discussion it is important to make a clear distinction between the publishing function and "The Publisher"as an institution. The last category can be a commercial publisher, a society publisher, a university press, an individual author or even the manager of a data base. In the same vein we have to make clear the distinction between the functions of the library (digital or traditional) and "The Library" as an institution fulfilling these functions.

Firstly, we will discuss the traditional publishing and library functions VI-1 –> VI-4, then we discuss the growing knitting together in VI-5.

VI-1. The **Traditional Publishing Functions** comprise the final stage of production of the comprehensive reporting of scientific results that the authors deem necessary to make publicly available. In general these are the more-or-less successfully terminated research endeavours (or parts thereof).
We can list the following functions:
a- The definition of instructions to authors in such a way that an unambiguous final presentation of the work becomes possible. This is a production driven function.
b- The organisation of the validation of the manuscript by peer review procedures. These procedures are subject of a permanent debate and vary per field and kind of certification.
c- The certification of the work by publishing it in a Journal. A journal can be best defined as a unique combination of subject field, quality level and genre. The journal title is the metadata that symbolises the certification, and is an important metadata in a retrieval environment.
The last two functions are both culturally driven as well as defined with respect to content. They are highly social processes in which prestige and personal relationships between the publishing staff and the editorial boards of the journals, who carry out the tasks, play a crucial role.
d- The physical production of journal issues.
e- The organisation of the subscribers administration.
f- The dissemination of the journal to the subscribers, including marketing of the journal and its contents.
g- As a consequence of this process, in the traditional situation a limited amount of metadata is generated, such as the bibliographic references and in some cases also controlled keywords.
In the traditional situation the keyword indexing was done by so-called Abstracting and Indexing services also known as secondary publishing. On top of that the library has its own, subject field defined, indexing and metadata systems.

VI-2. The **New Publishing Functions** entail functions that traditionally are known as Library Functions (see below), whilst others are induced by the new electronic media.
a- The conversion of the accepted, validated, work from a plethora of text processing systems into a standardised XML and PDF version.
b- The co-development of industry standards and the development of Document Type Definitions (DTD)
c- The development and exploitation of an electronic Repository with all articles published in the course of years.
d- The development of metadata systems as well as search engines to enable browsing and searching in this Repository.
e- The integration of all disciplines in one platform, enabling interdisciplinary usage.
f- The integration of non-text elements as part of or as an appendix to a work.
As a consequence of the electronic production the publication of single articles contrary to collections of articles in journal issues becomes the prevalent way of publishing. This fact makes the metadata <journalname> even more important as a symbol for field, quality and genre.

VI-3. The **Traditional Library Functions** entail the collecting, keeping and managing of those information sources that are needed for the contemporary research and educational interests of the institution. In that sense we have to make a distinction between the library

functions and the archival and museological functions. In essence the functions circle around the management, integration and disclosure of sources essential for research and education.
a- The collecting of own productions, often only master and doctorate theses.
b- The dedicated collecting of research, reference and educational sources in printed form, defined by the needs of the institute's staff.
c- Indexing the relevant material in bibliographic sense and according to subject.
d- Enabling access, also to other libraries and information sources, by cataloguing.
e- Educating students and staff how library systems work and assisting in finding relevant material.
f- Public functions like reading rooms, lending facilities and awareness services.

VI-4. The **New Library Functions** in the electronic era expand and deepen considerably the traditional functions.
a- The collection from the great variety of external sources can be better homogenised in an electronic catalogue.
b- The own collection can be more easily spearheaded to the direct needs of staff and students.
c- The website of a library develops to a client targeted portal of only relevant material for the user.
d- The collection of all works created by the Institute's staff and students can be easily integrated in the library system.
e- The access to external material world wide becomes trivial (provided the relevant technology and rights are covered).
f- The library develops to a multi-media knowledge communication centre.

VI-5. From the above it becomes clear that we have entered a transition period in which technology redefines part of the traditional organisational split in fulfilling the publishing and library functions. This split is one of the two reasons the Publishing House and the Library Organisation are two separated entities. The other reason is that culturally and sociologically, the craft of being a publisher is simply another craft than being a librarian.
We have left the times behind that publishing and printing were intertwined and located at one place, whilst the products needed labourious operations to ship them to libraries scattered around the world. Since about half a decade, the problem with electronic storage and logistics of works in an electronic representation is no longer the bottleneck, as long as we deal with more-or-less stable technological industry standards.
Driven by the possibilities to hyper-link works and to enable cross journal and cross article searching, the obvious step of the publishing industry (those organisations that commercially or not commercially fulfil the publishing functions) is to create virtual libraries of all the journal papers and books they ever published. These virtual libraries are still limited by the publishers' own production and are interlinked by metadata such as the Digital Object Identifier (DOI), by the clearing house Crossref. The aim and hope of the Publishing Houses are that with search technologies, they can avoid or overcome the librarian craft of dedicated selection, indexing and collecting.
On the other hand, by virtue of the same technology we see a development that libraries start to publish journals with the aim and hope to overcome and avoid the publisher's craft.
A new development is the open archive movement and the strong anti-publishing houses, self publishing movement.

Concluding it is important to realise that the publishers function is more than the organisation of the standard peer-review process and publishing houses (commercial and not-for-profit

alike) do play an important role in promoting not yet established science on new or even debatable issues. The free flow information, as history unfortunately proves, can be hampered by institutional interests and politics. In advanced, frontier or heretic fields standard peer-review often breaks down. The quality stamp must then be given by inherent coherence and not by agreed correctness. This aspect will certainly keep us busy as in an electronic environment quality control is much more difficult to define than in the past.

Part two
Do it yourself or delegate
The Institutional Repository as a new research platform

In order to reach a technological superior and economical responsible new system of creation, validation, certification, indexing, storing, archiving, disseminating, communicating, etc., etc. of the institution's intellectual output we have to make clear distinctions on what functions or tasks can be delegated or outsourced and what functions remain fundamental functions of the institutes of higher education. Such a new comprehensive system will transcend the traditional library and publishing organisations. A clear vision of this new situation is needed before it makes sense to compare prices and to discuss so-called business models.

I. The commercial and non-commercial aspects of cultural heritage

In our present cultural tradition and practice, all products of arts and science are considered to add to the total mass of cultural heritage for the advancement of humankind. Parts of these products are only of interest for the sake of knowledge, others play a role in the development of products and services, whilst for some items a private buyers market exists. Fundamentally there is also a division between *l'art pour l'art* and applicable results. In the sciences there is a strong drive towards useful research for society, though the claims and expectations are normally much higher than the practical results.
This immediately brings us to the thorny discussion of intellectual ownership, patents, copyright, trademarks and all the other Intellectual Property Rights (IPR).
As soon as a Repository, as described in part 1, section II, is in the making, IPR immediately pop-up. We can paint various scenarios:
- The IPR belongs to the employer. In that case it is the employer who can decide what to do.
- The IPR belongs to the creator. In that case the Institutional Repository needs a license to host the work. Otherwise a bibliographic reference link to a place where the work resides is the only possibility.
A direct consequence of this is that: filling the Institutional Repository rest either on the benevolence of the creator (and creators might leave the institution with a quarrel) or on a contract of the Institute with the creator which stipulates that all works made during his/her stay automatically become available in the Institutional Repository.

The picture is complicated by the fact that: 1) Books as mentioned in section II-9 can be in the realm of the general public, and 2) the publishing houses (commercial and society publishers alike) argue that copyright transfer of manuscripts (nowadays compuscripts) is compulsory to make their operations economically viable and to be able to fight against plagiarism.

Interestingly, this last argument is typical for the trade in single items. In this traditional view the document is like a painting or a sculpture of an artist. In an electronic environment where copies of electronic representations of any intellectual expression spread like a viral disease, this reasoning breaks down. It is for that reason that the intelligent publishing houses shift their argument towards the notion of information services provision. By doing so they incorporate part of the library functions.

Another important observation is that, at present, publishing houses are not very interested in

data-sets or other large digital object collections. Exceptions are the large proprietary data bases of, e.g., the pharmacological industry or the military.

The interest of commercial parties in exploiting content is to what extent they are able to develop a business model that at least covers costs.

Not all kinds of knowledge representation as mentioned in section II are fit to become commodities, but all of them are sources for the creation of commodities.

An important and promising development is the discussion on creative commons, know by the catch phrase *From all rights reserved to some rights reserved*. Here creators (authors) define themselves to what extent the IPR allow reuse in various forms and under what conditions (commercial or not). In a way this grassroots movement defending the free flow of information, tries to find a golden mean between a full scale IPR regime such as defended by the World Trade Organisation, in line with the present neo-liberal market economy ideology and a gentle policy that enables a creator to keep some responsibilities and rights without abolishing intellectual property rights as such.

II. Costs calculations and who is paying what

II-1. Despite many heated contributions in the discussions, no reliable cost calculations are available. An important reason for that is the rapidly changing prices in hardware and telecommunications and the impressive growth of software developments. It is obvious that the large publishing houses could keep their economical performance by sharply reducing costs, such as the reduction of their copy edit efforts, the strong emphasis on automated logistics and production processes, and the export of all typesetting labour to India, Singapore and China. This standard market economy throat-cutting practice allows us to safely assume that for journal articles and books the sheer production costs of works in XML and PDF are the lowest possible for these mammoth companies. This observation does not say anything about the remainder of their costs edifice, editorial costs and inter-company costs, which determines their sales prices. The prices mentioned by publishers, Open Archive aficionados or consultants are ridden with uncertainties, as no clear cost price calculations are given in relation to the: technical complexity of the manuscripts, output quality and standardisation, editorial enhancements, overheads, etc. Also often sales prices are mentioned without a clear insight in the level of (hidden) subsidies, profit marges, tax levels, etc., etc.

II-2. In many discussions it is mentioned that the institutions pay in every phase. They first pay the research, subsequently they pay the editorial work and peer reviewing and then they pay the subscriptions. Rhetorically this sounds as a strong argument, but it is useful to follow the value chain a bit closer.
a- It is obvious that the research and educational costs of those works that might fill an Institutional Repository are paid by public funds and, in some cases, commercial grants.
b- With the advent of the personal computer, the typing of the manuscript is no longer done by low paid administrative staff, but by the scholarly staff themselves. In a way, part of cheap labour is converted to more expensive labour. Most presumably, this development is here to stay and the full costs of the process can only be reduced if this expensive labour is not only used for typing but also for, e.g., indexing on the fly, and other enrichment activities that demand domain knowledge. R&D on dedicated author environments will address this issue

directly. That way the labour costs remain high, but we get a higher quality and more versatile product.
c- Editing and refereeing are part of the academic pursuit. It is the codified process of mutual critical self-analysis and discussion. It obviously belongs to the academic craft. As long as we consider the results of academic research a good that belongs to society, editing and peer review belong to that societal good. In the case of theses, working papers, reports, etc., the organisation is simply part of the job.
However, the organisation of this process for journal articles is a special non-academic craft that is delegated to publishing houses. Due to the economy of scale and nowadays the highly computerised logistics, the cost price is lowering.
d- The conversion in an accepted standard according to a well-formed structure, e.g., an XML-DTD is certainly not the task of the scholar. As said above, many publishing houses make use of the services of Asiatic specialist companies. Hence, there is no reason to reinvent the wheel and an Institutional Repository must outsource that task to the best buy. It goes without saying that this pertains for all textual items in an Institutional Repository and not only for preprints and journal articles.
An important issue is here that such a standardisation belongs to the data level and not to the service level.
e- The storage and distribution costs, provided that paper copies are only produced by the reader at his/her own expense (which can be high in the cases that colour is essential), are low and some suggest that they can be piggy bagged on the academic infrastructure. This might be the result of a full cost price calculation at the end of the day.
But prior to that we need a good understanding of the full costs. We have clear evidence that these baneful commercial publishing houses spend hundreds of millions of Euros in building their electronic warehouses with all logistics, bells and whistles. The most advanced operation of Elsevier processes about 1000 scientific articles per day. Given this economy of scale, it is difficult to argue on pure financial grounds that a distributed Institutional Repository operation can do the same job cheaper (again without the inter-company overheads and profits).
It is another question to what extent the buyer is able to purchase (and/or get access to) part of the material in the repository or that only the full product is available, as is card played by the Commercial Publishers. To cover the costs of the whole Repository, schemes have to be worked out. Many items stay put and wait (see part1, I-5), other items might know a short or long period of popularity. In a non-commercial approach one can argue that the repository is just an academic asset that belongs to the infrastructure and its existence is fully warranted by the new academic output, based on the Repository.
f- The marketing and sales efforts of publishing houses can of course diminish as soon as all academic material becomes freely available. However, in that case we still have three caveats:
1) Do we, and if yes how, charge commercial users?
2) In such a system, the whole financial picture goes topsy-turvy as at present the costs are covered by the library budget. A completely new cost allocation system has to be worked out. In dealing with scholarly journal publications only, we already have a heated debate what is cheaper: an Open Archive (OA) approach or the present subscription model. The known parameters are whether the institute is a net reader or a net writer of papers. In an OA case the advantages of the subscription model are gone. In a subscription model immediately needed information and contextually related information are delivered in one go, therewith averaging out the total production costs over all products. The present OA examples are limited in size and no calculations on scaling to millions of items are available.
3) Awareness services and search methodologies remain crucial to disclose and retrieve relevant information.
The conclusion of this section is that we certainly can try to calculate the various cost factors

in the value chain, but that we also have to be very certain about precisely what (sub)tasks we want to keep as an institutional activity and which one we see as part and parcel of the academic intellectual pursuit. We have to make clear choices what to out source and what not. In the present discussion the clarification of the question to what extent an Institutional Repository adds to the primary goals of the various academic disciplines is the most important short term task.
In that sense the continuing changing financial figures are less interesting than the understanding of the changing process of knowledge representation and management.

III. The Institutional Repository is more than a weapon in the battlefield of journal publishing

In the above we have tried to delineate the various factors and issues in the encompassing topic of Institutional Repositories. Starting with the mission of the institutes of higher education, we worked through the various tasks and factors that determine an Institutional Repository. The goal of this exercise is to make clear what type of content an Institutional Repository harbours and to what extent the devolvement of an Institutional Repository is currently in competition with other parties in the information market. From the various, but still limited studies on price competition in the journal market between an open archive environment and professional publishing enterprises, we learn that the key factor in an author paid system (the equivalent of the very popular system of page charges mainly US society publishes adhered to till about the seventies of the previous century) is the type of institution. Is it an institution in a field with a high publication rate such as synthetic organic chemistry or a field in which a single publication demands a lot of reading, such as literature studies.

Given the current emphasis on journal articles, it is an obvious reflex of the community to concentrate on this type of scientific works. But it is also here that we know the best oiled operations. The argument that these operations are often run by stock market driven companies may not determine the whole central issue of Institutional Repositories, namely that they are the ultimate representation of all the scientific and educational output of an institution.

In an electronic environment that will lead us to genuine e-science, that is to say science that is fully determined by a digital environment, journal articles will more than ever only be the final phase of research endeavours. A full stop after the fact. In some fields, the current discussion is already fully based on preprints and the final article only serves the role of certification for administrative purposes. In those fields the published articles are almost solely for the non-reader as described in part 1, IV-1.

The central question is not to what extent do we build Institutional Repositories to battle with the publishing houses. In our market economy their concentrated force is big and the academic institutions are under competitive stress. This does not mean that we have to join them if we can't beat them. The struggle on (sales) price structures, services etc. will go on and will be augmented by serious cost price calculations for an internationally linked system of Institutional Repositories, still to be made.
But having said that, the essential discussion is not the standardised data level, including converting materials to an international standard, securing access and guaranteeing digital preservation. The essential discussion is to what extent the Institutional Repositories fulfills a

service role in academic life. This role can only become successful if all academic stakeholders realise that a Repository is not a library in new cloth.

The novel thing in the whole story is that a genuine academic Institutional Repository, which is transparently linked to others, creates a totally new platform for academic activities. By integrating all research and educational endeavours in one transparent digital infrastructure, the Institutional Repository becomes a research tool in itself.
In the past it has been frequently defended that the library is the laboratory for the humanist. In an electronic environment, where all kinds of information, including primary sources and raw results are intertwined, this laboratory function will become obvious for all disciplines.

In academic life the Repository will soon play the role of the central metabolic organ for knowledge. In that sense it has to be rated on the same level as buildings, fresh air and water, all needed to sustain a healthy body.