OAI, Google Scholar and Wikipedia are the answers, but what is the question?

Marlon Domingus
Projectmanager Open Access Leiden
Leiden University Library

OAI4 – CERN
21 October 2005

Universiteit Leiden

Route

1. Wikipedia: contours of the theme
2. Collaboratory: science community-wise
3. Some observations
4. Science vs Scholarship: a first glance
5. Basic armchair philosophy
6. New directions for reseach

Wikipedia

Assumption 1: the wiki-technology supports a new way of creating knowledge on the web.

Wikipedia: talking numbers

Wikipedia.org currently has a total of 776,068 articles and 220 million words.

The encyclopedia is 4 years old
(started in January 2001)

Encyclopædia Britannica online, has an estimated 120,000 articles and 77 million words
(December 2004)

Wikipedia has 4.5 times the number of articles and nearly 2.5 times as many words.

Source: http://en.wikipedia.org/wiki/Wikipedia:Largest_encyclopedia

Universiteit Leiden

Wikipedia: key features

Quantity (breadth and depth)

[total of 776,068 articles and 220 million words.]

Speed

[The encyclopedia is 4 years old]

Quantity (breadth and depth)

[Wikipedia has 4.5 times the number of articles and nearly 2.5 times as many words.]

Universiteit Leiden

Wikipedia: quality?

Robert McHenry [former editor in chief, Encyclopædia Britannica] on Wikipedia:

1. Anyone can submit an article and it will be published.
2. Anyone can edit that article, and the modifications will stand until further modified.

Then comes the crucial and entirely faith-based step:

3. Some unspecified quasi-Darwinian process will assure that (...) articles will eventually reach a steady state that corresponds to the highest degree of accuracy.

Source: R. McHenry, The Faith-Based Encyclopedia. www.techcentralstation.com, 2004

Universiteit Leiden

Wikipedia: quality?

McHenry basically holds / implies that

1. in Wikipedia.org reviewing doesn't work

2. because articles are debated and edited continually by everyone

So McHenry must accept that quality can be achieved within Wikipedia.org once these issues are dealt with. Currently, this is the case.

Universiteit Leiden

Wikipedia revisited

Wikipedia.org today:

1. Does indeed consist of relevant articles of respectable quality and size, at least compared to Encarta Pro and Brockhaus. Source
2. Articles do get regular reviewing and closing an article is considered once the experts have agreed on an article*
3. Non trivial policies for starting and editing an article are implemented*
4. Does, however, lack a number of important articles, still waiting for volunteer netizens to be started*

* Source: H. Blankesteijn, *Wie het weet mag het schrijven. Zelfcorrecties geven internet-encyclopedie Wikipedia kwaliteit.* NRC, pg. 1 W&O Bijlage, 19/20 februari 2005.

Universiteit Leiden

Wikipedia revisited

The real trick of Wikipedia though, is:

1. The *collaborative* aspect, within the Wikipedia community, which accelerated the pace of articles published (speed and quantity).

> Wiki software must be the most promiscuous form of publishing there is--Wikipedia will take anything from anybody.*

* Larry Sanger, Wikipedia is wide open. Why is it growing so fast? Why isn't it full of nonsense? (Op-Ed) 2001

Wikipedia revisited

2. The social aspect of the *collaborative* way.

> We (that is, we participants) work on a lot of different pages, and I think most of us feel some **collective responsibility** for how the whole thing looks. We're constantly cleaning up after each other and new people.
>
> In the process, a **camaraderie--a politeness and congeniality** not found on many online discussion forums--has developed. We've got to **respect** each other, because we are each other's editors, and we all have more or less the same goal: to create a huge, high-quality free encyclopedia.
>
> The way I see it, we're having fun creating a thing of beauty.

* Larry Sanger, Wikipedia is wide open. Why is it growing so fast? Why isn't it full of nonsense? (Op-Ed) 2001

Universiteit Leiden

Driver for Wikipedia

What drives the volunteer netizens / Wikipedians / Wikipediholics to publish articles and to maintain the quality of 'their' articles in Wikipedia.org, all for free?

Assumption 2: the principle of the gift exchange society*:

'The act of publication constitutes a gift exchange among a community of devotees bound by a common interest; the giving of such gifts is intended to win the regard of other members in the community.'

* Source: To Publish and Perish. Policy Perspectives, 1998

Pre-requisites for Wikipedia

What are pre-requisites for these volunteer netizens / Wikipedians / Wikipediholics to publish articles and to maintain the quality of 'their' articles in Wikipedia.org, all for free?

1. collaborative building
2. voluntary membership of a community
3. easy access and ease of use of tools

Wikipedia: free knowledge for all

Wikipedia.org proves that people will
succesfully volunteer to collaborate
in building a free online encyclopedia
of some quality and this quality is maintained
by selfregulation within the community,
given easy access and ease of use of tools.

From Wikipedia to science

Assumption 3: the pre-requisites:

1: collaborative building
2: voluntary membership of a community
3: easy access and ease of use of tools

also work for science and are, combined, present in the concept of a “Collaboratory”.

Assumption 4: these pre-requisites themselves are not new for science

Universiteit Leiden

Collaboratory - when

In 1988 W.A. Wulf [Director of the National Science Foundation's (NSF) Directorate for Computer and Information Science and Engineering (CISE)] introduced the word: "collaboratory" as a blend of the words collaboration and laboratory.

Source: The National Collaboratory – A White Paper (Wulf, 1988)

Note to the famous Wulf White paper

The Wulf White Paper was first identified as cited in the Lederberg and Uncapher report in a footnote to one of the retrieval set documents. But, neither document was cataloged as held by any lending library in the world, nor were they indexed as available in any database. Neither was either available directly from the National Science Foundation, or from the authors.

The Lederberg and Uncapher report, which contains the White Paper as an appendix, was eventually tracked down by an enterprisinging access librarian at TWU, Joe Natale, who called upon a librarian at Rockerfeller University, who gladly descended into the bowels of the library, located a dusty box of documents remaining from the 1989 workshop, and thumbed through the entire contents until she located the report, which she duplicated and delivered.

Source: footnote 1, pg 15, Joanne Twining, A naturalistic journey into the collaboratory: in search of understanding for prospective participants. Dissertation, Denton, Texas, 1999.

Universiteit Leiden

Collaboratory – a concept explained

Wulf describes* his concept of a Collaboratory:

"a center without walls, in which the nation's researchers can perform their research without regard to physical location – interacting with colleagues, accessing instrumentation, sharing data and computational resources, [and] accessing information in digital libraries"

* Wulf, *The Collaboratory Opportunity*. Science, Vol. 216, 1993

Collaboratory / Telescience

Peter Bank [Dean, college of Engineering, University of Michigan] reacts to Wulf's 'Collaboratory Opportunity' by stating that the collaboratory principle was already demonstrated by NASA in 1980 (by operating the International Ultraviolet Explorer, a major scientific satellite), then called: "Telescience".

Bank agrees however that "Labels aside, such types of electronic group interactions are growing rapidly and offer more profitable means as bulletin boards and electronic mail".

* Bank, *"Collaboratory" Principles*. Science, Vol. …, 1993

Universiteit Leiden

Collaboratory: a rationale

Major research centers and laboratories all over the world provide environments that have accelerated the pace of scientific advances.

As the frontiers of knowledge are pushed back the problems get more and more difficult, often requiring large, complex teams – frequently multidisciplinary – to make progress.

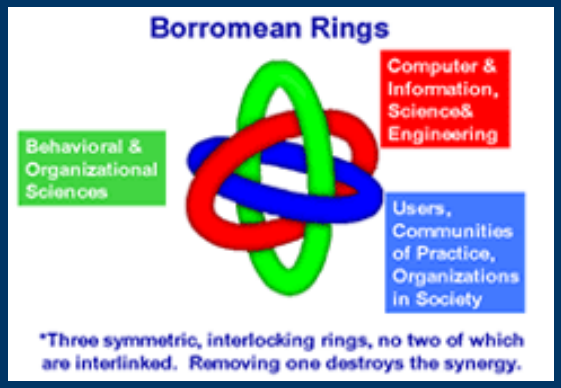
Exotic and expensive equipment or facilities can only be justified if they are designed and deployed collaboratively, maximizing their impact on a science community.

Source: p. 44, Gary M. Olson, et al., *Collaboratories to Support Distributed Science: The Example of International HIV/AIDS Research*. Proceedings of SAICSIT 2002.

Universiteit Leiden

Collaboratories: community-driven

The Science of Collaboratories project is devoted to understanding the technical and behavioral principles that can lead to better, more successful design of collaboratories in the future.



The project is being driven by researchers at the University of Michigan, and Howard University, and is sponsored by the National Science Foundation.

community-membership

Collaboratory: easy access and ease of use

An increasing number of scientific problems call for or would benefit from collaboration among researchers, while improvements in the capabilities, ease of use, and availability of computing and communications systems suggest that information technology can facilitate and enable collaboration.

Source: preface, National Collaboratories: Applying Information Technology for Scientific Research. National Academy Press, Washington, 1993.

Collaboratory as community-driven science

A collaboratory proves that science can be done community-wise and by combining resources and technique a new corpus of knowledge of high quality emerges. This quality is maintained by selfregulation within the community, given easy access and ease of use of tools.

Recommendation: regard science as communities of practice, a phenomenon from the situated learning environment.

Some observations

Within the project Open Access Leiden some moderate experience with adoptation and implementation of OAI-based data-infrastructure and basic OAI-services.

Focus currently on Humanities, Law and Archeology.

A few things become clear pretty soon:

1. Different communities of researchers perceive OAI differently and see specific opportunities for them

2. The level in which research will indeed be supported by OAI / IT depends in the level of understanding of the way research works

3. Focus on the research and the communities of researchers. “Find the differences.”

Universiteit Leiden

Science vs Scholarship

The distinction science / scholarship is only helpful in part: it helps to cluster the images we have on the supposedly quite different modi operandi.

The most obvious remark one can make about this distinction with regard to OAI is that in the Sciences researchers as a rule already have made OAI work for them.

Within the ´Scholarship´ type of reserach OAI is known at best or still a buzz word.

Epistemological conflict: certain branches of research within Archeology do not fit in neatly within the science / scholarship distinction.

Focus on research - a first glance: indicators for the way research is done

1. Publication type (monograph – article – proceeding)
2. Peer review
3. Amount of publications a year
4. Size of a typical publication
5. The amount of changes in content done by third parties after submitting a pre-print
6. The number of co-authors of a typical publication
7. The relevance of a publication in years
8. What is the typical medium for a publication
9. What are the discipline based publishing traditions
10. Are citations to the own publication of importance
11. Are the journals in which one publishes ranked
12. What is the attitude towards open access
13. What is the policy with regards to exclusively transfering rights to publishers.
14. What method is used to further research (experiment / debate / …)

Universiteit Leiden

Focus on the researcher: a first glance

What then are indicators for the way the researcher works?

1. Is the discipline by nature conservative or innovative
2. Do the debates get closed or as a rule not
3. Does the discipline create own technical solutions
4. Does the discipline entail modelling / rendering / imaging / computation / simulation
5. What kind of data is the research based on
6. What is the amount of data the research based on
7. Is disciplinary-specific IT in use
8. Is research a way to get funding

Universiteit Leiden

Differences: the typical scientist

1. Writes mainly articles, proceedings and sometimes a monograph
- Peer review is important
- A fairly large amount of publications a year
- A typical publication is has > 10 pages
- The amount of changes in content done by third parties after submitting a pre-print is small
- The number of co-authors of a typical publication is large
- The relevance of a publication in years is small
- The typical medium for a publication is digital
- Publishing tradition: fast and collaborative
- Citations to the own publication are of importance
- The journals are ranked
- The attitude towards open access: positive / no issue
- Policy with regards to exclusively transfering rights to publishers: no issue due to focus on pre –prints
- The experiment is used as a method to further research

Universiteit Leiden

Differences : the typical scholar

1. Writes first and foremost monographs
1. Peer review is indirectly important
2. A fairly small amount of publications a year
3. A typical publication is has > 75 pages
4. The amount of changes in content done by third parties after submitting a pre-print is large
5. The number of co-authors of a typical publication is 0 to 1
6. The relevance of a publication in years is large
7. The typical medium for a publication is paper
8. Publishing tradition: slow and solistic
9. Citations to the own publication are of importance
10. The publishers / journals are ranked
11. The attitude towards open access: reluctant / an issue
12. Policy with regards to exclusively transfering rights to publishers: an issue due to focus on post prints
13. Debate is used as a method to further research

Universiteit Leiden

Making most of differences

OAI-services for a typical scientist

1. Will entail Web of Science / Web of Knowledge
2. Will be based on pre prints
3. Will focus on publications and datasets
4. Will be innovative; using OAI to do new things
5. Will probably exist already
6. Will be low cost, for based on existing products – infrastructure
7. …

Universiteit Leiden

Making most of differences

OAI-services for a typical scholar

1. Will sometimes entail Web of Science / Web of Knowledge
2. Will be based on post print publications
3. Will have rights issues
4. Will be traditional, doing the same things but now with OAI
5. Will be expensive, for based on retro scanning and retro activities
6. ...

Classification issue: in Law citations are non existent [no WoS and no Google Scholar needed], journals are not ranked, articles are not refereed.

Science in action vs ready made science

A convenient way to make a distinction between science and scholarship is to use the distinction Bruno Latour formulated in his: *Science in action* (1987):

Science in action vs ready made science

In Science the debate is based on the pre prints: the phase where science is being made.

Latour examines the sociological and rhetorical principles in this phase and illustrates what makes some theories become succesful and others not.

From an OAI-perspective: pre print services will be usefull for science for it has the main focus. Indeed these services usually already exists in some form.

Science in action vs ready made science

Once science is ‘made’, following Latour, it becomes ‘ready made science’; is accepted and it gets printed in the text books as part of the intellectual corpus of the discipline. There is stayes relatively fixed and becomes a black box.

The real debate however stays within the science in the making.

From an OAI-perspective: a post prints service will be marginally usefull for science. Perhaps useful for:

1. visibility and access
2. making the citations visible and
3. printing on demand services
4. archiving

Universiteit Leiden

Research after the first glance

How to proceed to get more specific and accurate insight into the way research works?

Research at our Law department as well as our Archeology Department don't fit neither of the two presented 'typical' research-styles.

Strategy 1: let's look at what philosophy of science tells us to break down the question.

Four views

1. Philosophy of science is the formulation of world-views that are consistent with, and in some sense based upon, important scientific theories.

2. Philosophy of science is an exposition of the presuppositions and presuppositions of scientists.

3. Philosophy of science is a discipline in which the concepts and theories of the sciences are analysed and clarified.

Source: Introduction, John Losee, *A historical introduction to the Philosophy of science*. Oxford, 1980

Four views

4. Philosophy of science is a second-order criteriology in which questions are raised such as:

(a) what characteristics distinguish scientific inquiry from other types of investigation?

(b) What procedures should scientists follow in investigating nature?

(c) What conditions must be satisfied for a scientific explanation to be correct?

(d) What is the cognitive status of scientific laws and principles

Universiteit Leiden

Four views

The fourth view seems to be the most suitable for our questions

(a) what characteristics distinguish scientific inquiry from other types of investigation?

This refers to the type of questions singled out in the list of indicators for the way research is done. This list obviously needs some more work though.

In Leiden we are currently testing and fine-tuning this list. Results will be shared of course.

What you probably will get is specific domains within the ‘Science / Scholarship’-scale that can be clustered based upon specific characteristics they share.

Four views

(b) What procedures should research follow in investigating nature?

Scientific method: does it exist or not? What are the characteristics?

1. Inductivism
2. Hypothetico-deductivism
3. Mathematical positivism
4. Conventionalism
5. …

Four views

(c) What conditions must be satisfied for an explanation to be correct?

1. experiments vs debate
2. …

Four views: cognitive status of scientific laws and principles

(d) What is the cognitive status of scientific laws and principles

This refers to one of the points Latour made in his Science in action study. It gave us the distinction of pre print services vs post print services.

Four views: cognitive status of scientific laws and principles

See also a logical reconstructivist approach of philosophy of science:

Level	Content	Example
Theories	Deductive systems	Kinetic molecular theory
Laws	Invariant relations	Boyle's Law
Values of concepts	Statements that assign values	V = 1,5 lit.
Primary expermental data	Statements about pointer readings etc.	Pointer B is on 3.5

Source: p 175, John Losee, *A historical introduction to the Philosophy of science*. Oxford, 1980

However

Science and Scholarship are ever changing:

the scientific method
the type of questions
the amount of computing
the way research is funded
the way research is organized

More in general: what will the concept of E-Science bring about in the way research is done, how will collaboratories, and how will the Grid?

Also: how can we as the OAI-community help change* research?

* See the work done in: Herbert Van de Sompel, Sandy Payette, John Ericson, Carl Lagoze, Simeon Warner, Rethinking Scholarly Communication: Building the System that Scholars Deserve. D-Lib Magazine, September 2004.

Scholarship and science

In Scholarship we find:

1. the *scientific* method is becoming more of a standard (research is funded if research is done in research schools; citations are important as is publishing in high ranked journals.)

Constantinos Dallas urges scholars in this respect to adapt strategies from the sciences, for instance to strategically collaborate more and thus become a stronger, organized discipline instead of competing each other within the discipline in favour of other disciplines.

2. the type of questions are changing – due to larger research groups, topics are becoming less specific and less interesting for researchers

* Source: Constantinos Dallas, Humanistic research, information resources and electronic communication. 1998.

Universiteit Leiden

Scholarship and science

3. the amount of computing - see Humanities computing*:

- datastructures: (historical images, texts and grammatical structures)
- dataprocesses and algorithmic
- interface development

Source: Adriaan H. van der Weel en Gerhard Jan Nauta, *ICT-nota Letterenfaculteit; deel "Humanities Computing" (onderwijs & onderzoek)*. Leiden, werkversie december 2000.

Universiteit Leiden

Scholarship and science

4. the way research is funded

Funding is more and more based upon principles accepted within the sciences (many citations and publications in ranked journals give access to a larger portion of the cake).

In general: decreasing direct and increasing indirect government funding, emphasis on social relevance, a policy of improving quality and further internationalisation.*

5. the way research is organized

More and more large (inter)national research schools have evolved.

* Source: the forthcoming article by C.J.J.M. Stolker: *Legal journals: in pursuit of a more scientific approach*.

Additional questions

Strategy 2: the subject of research.

One could also look at the *subject* of research and narrow down differences between scientists and scholars and formulate services.

In Humanities Computing, Humanities are understood to have the following characteristics:

1. a narrative, textual approach is adhered
2. there's no one definitive insight: different and conflicting views are valuable and part of the discours
3. the number of elements of a complex object (painting / text) is innumerous; the material is rich and it can hardly be reduced

Universiteit Leiden

Additional questions

Strategy 2: the subject of research.

Law is understood to have the following characteristics:

1. legal journals are in pursuit of a more scientific approach*, for:

> Two things are wrong with almost all legal writing. One is it's style. The other is its content.

2. strong national focus
3. individualistic nature
4. publishing culture: normative, commentative, lacking an explicitly defined scholarly method, with little interest in empirical research

* Source: the forthcoming article by C.J.J.M. Stolker: *Legal journals: in pursuit of a more scientific approach*.

Universiteit Leiden

And so on…

With some basic armchair philosophy, thinking about relevant principles and structures within science and scholarship, there's many places where the OAI-domain is relevant.

OAI is succesful because it is fairly simple* to set up an IR and develop services. The technical part is not so much the challenge as is the translation of the different aspects of research into support by OAI-services.

* The term "light-weight protocol" [Eric Morgan, *Using OAI-PMH and other light-weight protocols to facilitate scholarly communication*. CERN, 20 October 2005] is instructive here.

Universiteit Leiden

Scholarship next?

hypothesis 1:
the pre-requisites:

1: collaborative building
2: voluntary membership of a community
3: easy access and ease of use of tools

can work similarly for scholarship and the impact could potentially be ‘larger and more profound’* if computing is done to do unconventional (new) things [E-Scholarship], not merely the same things with the use of IT [E-supported Scholarship].

* Source: *The Nature of Engineering, the Science of Humanities, and Godel's Theorem. Talking with William A. Wulf, President of the National Academy of Engineering*. Ubiquity.

Universiteit Leiden

New Directions for reseach. OAI based?

1. Research outside universities in research schools

fact: three of the last four Spinoza laureates are rewarded for their research in their research schools – not their universities

2. Post Humboldt universities?

Hans Wissema, *Technostarters en de Derde Generatie Universiteit*. Oratie, Delft, 6 juni 2005