

Access IT Training

**What's next?**

---

# Yesterday

- 2003
  - Google indexed 3,3 billion of pages
    - <http://searchenginewatch.com/3071371>
- 2005
  - Google's index contains 8,1 billion of websites
    - <http://blog.searchenginewatch.com/050517-075657>
  - Estimated size of whole searchable internet - 11,5 billion of pages
- 2010
  - No one counts this anymore

# Yesterday

- 2001
  - First Open Archives Initiative workshop at CERN
    - <http://indico.cern.ch/conferenceDisplay.py?confId=ao1193>
  - The first version of OAI-PMH specification
- 2002
  - OAI-PMH 2.0 specification was released
  - The dawn of OAIster.org
    - 66 repositories with 235 116 records
- 2005
  - OAIster.org has more than 5mln of records (400 repositories)

# Yesterday

- May 2008
  - OAster.org goes beyond 1000 of repositories with 1.5 million of search hit monthly
- October 2008
  - OAI-ORE spec. was released
- November 2008
  - Europeana prototype was launched (and crashed)
  - Initially it contained 2mln of objects.
- December 2009
  - Europeana reaches 5mln of objects
    - Polish resources are now available in Europeana!

# Agenda

- Semantic web
  - Where is the benefit?
  - Knowledge organization systems in SW
  - Web of data
- Transparent science
  - Workflows, datasets, articles
  - OAI-ORE

# Agenda

- Who will do all those things?
  - Crowdsourcing
  - Community collection building
- Conclusions

# Better tools

- Do you remember a world before YouTube, Flickr, Google Maps, Gmail, Wikipedia?
- “The Network is the computer” is a fact
- Better tools allows to create more content
- New media brings new challenges
  - Preservation of information stored in social portals like FaceBook, LinkedIn
  - What will happen to Second Life in 10/20 years?

# Semantic web



- Semantic web is supposed to extend capabilities of WWW
- How it will be done?
- What the semantic web is about?

- **What the semantic web is about?**
  - At the moment websites are designed for humans
  - SW is a vision of information that is understandable by computers
  - Thanks to this machines would be capable to perform more complicated tasks

- The concept of SW comprises a set of design principles and a variety of enabling technologies
- Technological foundation of SW relies on **Resource Description Framework (RDF)**
- RDF is a data model
- It is based upon the idea of **making statements about resources** in the form of subject-predicate-object expressions

- These expressions are known as *triples* in RDF terminology
- Subject denotes the resource
- Predicate denotes traits or aspects of the resource
- It expresses a relationship between the subject and the object

- RDF can be expressed in various serialization formats (including XML)

# Semantic web



- “The sky is blue”
  - Subject: the sky
  - Predicate: “has the color”
  - Object: blue
- Subject of an RDF statement is URI or a blank node e.g.
  - <http://dl.psnc.pl/biblioteka/dlibra/rdf.xml?type=e&id=207>

- Another element of SW technical side is Web Ontology Language (OWL) and RDF Schema
- OWL deals with a formal description of **concepts, terms and relationships** within a given **knowledge domain**
- OWL is a family of knowledge representation languages for authoring ontologies
- OWL ontologies are usually written as RDF/XML files

- e.g., an ontology describes families
  - It include predicates like "hasMother", "hasParent"
  - Individuals of class "HasTypeOBlood" are never related via "hasParent" to members of the "HasTypeABBlood" class
  - Thanks to those information some things can be derived from data

- Imagine an individual named Adam who is related via "hasMother" to individual Jolanta
- Adam is also a member of class HasTypeOBlood
- Thanks to these information it can **inferred** that Jolanta **is not** a member of HasTypeABBlood

# RDF Schema



- RDF Schema is also an extensible knowledge representation language
- It is less expressive than OWL

- Simple Knowledge Organization Systems is a family of formal languages designed for representation of :
  - Thesauri
  - Classification schemes
  - Taxonomies
  - Subject-heading systems
  - Any other type of structured vocabulary
- SKOS is built upon RDF and RDF Schema

- Its main objective is to enable easy publication of controlled structured vocabularies for Semantic Web
- Some important vocabularies are already available in SKOS format, e.g.
  - Library of Congress Subject Heading (LCSH)

- OWL is intended to express complex conceptual structures, which can be used to generate rich metadata and support inference tools
- SKOS is a simpler format, it can be extended to OWL

# Semantic web criticism



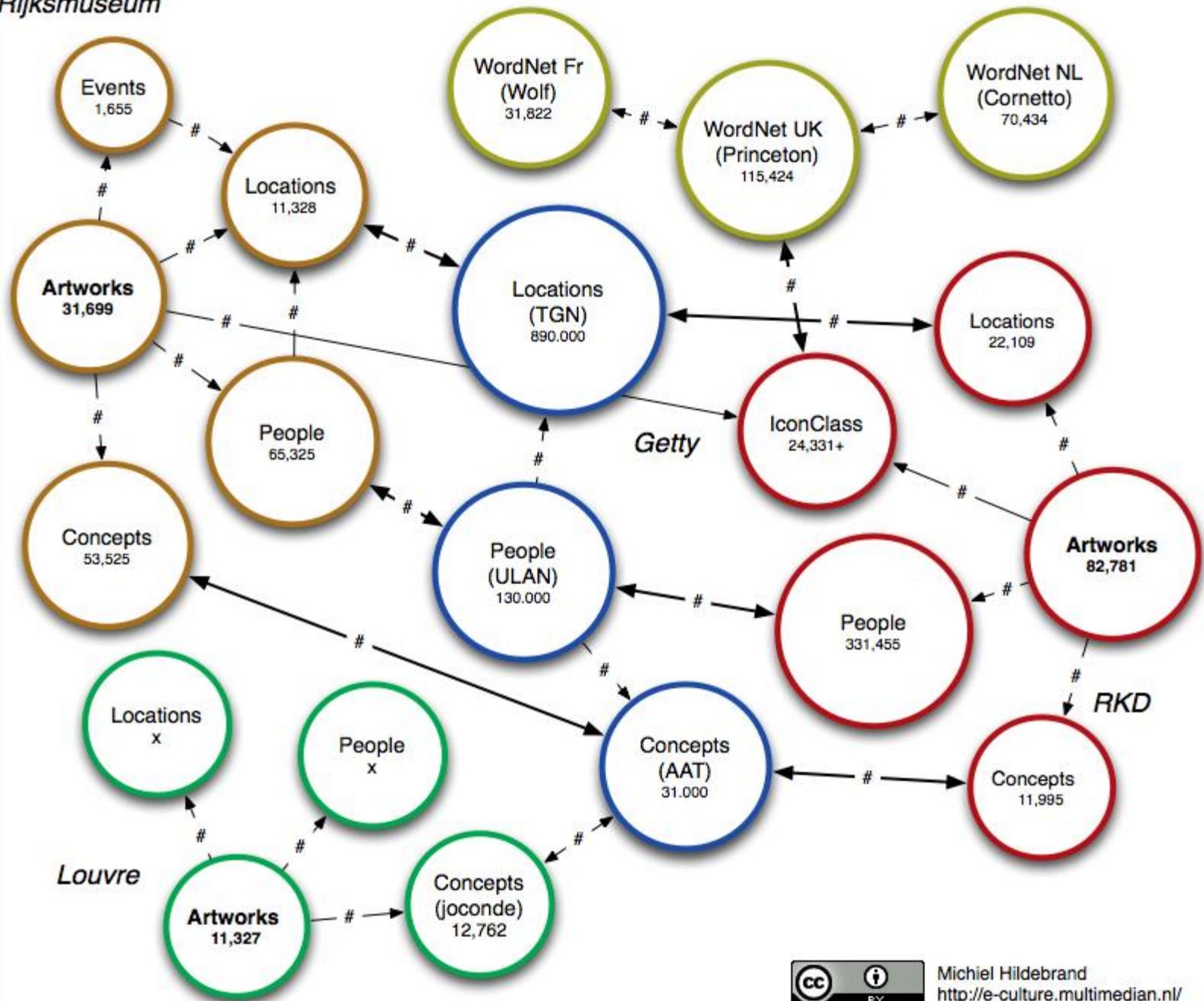
- Critics question the basic feasibility of a complete or even partial fulfillment of the semantic web
- Large scale utilization raises a lot of issues
  - Which ontology is the right one?
  - Who will create all those ontologies?
  - Who will prepare descriptions for web resources?

# Semantic web criticism

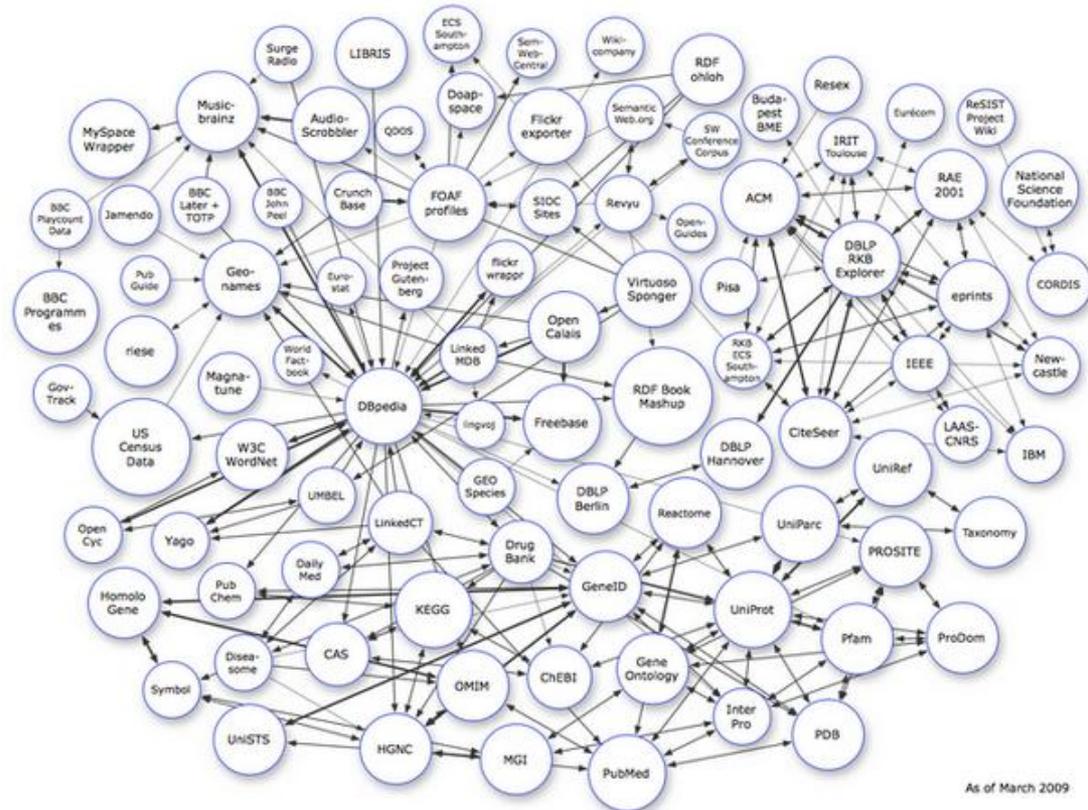


- There are also too few tools which support creation of semantic web resources
  - Some people are saying that there is no killer-app which will convince people to use semantic web

- SW offers a set of techniques which can be practically adopted in particular domains
- We are not looking for universal solution for all problems of the world, e.g.
  - multilinguality in the domain of cultural heritage
  - Europeana Thought Lab :  
<http://www.europeana.eu/portal/thought-lab.html>



- Linking Open Data (LOD) a W3 Consortium project



- Linking Open Data (LOD) a W3 Consortium project
- It attempts to connect various freely available data sets
- Data sets are set up to re-use existing ontologies such as WordNet, FOAF and SKOS to interconnect them
- LOD currently counts more than 2 billion of RDF triples

- Participating data sets:
  - <http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets>
- Example:
  - <http://dbpedia.org/page/Veria>

# References



- Wikipedia :
  - Semantic Web, OWL, SKOS, RDF, RDF Schema
- W3C Semantic Web Activity
  - <http://www.w3.org/2001/sw/>

# Transparent Science

- What is a results of scientific studies?
  - Paper in a journal
  - PDF at author's website
- Does it provide enough information to repeat given experiment and verify result on your own?
- What about dataset? What about workflow?

# Transparent Science

- Paper should be followed by additional resources
  - Workflow documentation
    - My experiment - is a collaborative environment where scientists can safely publish their workflows and experiment plans, share them with groups and find those of others
      - <http://www.myexperiment.org/>
  - Dataset used during the experiment
- How to associate those item ?

# OAI-ORE



- This is why OAI-ORE was initially developed
- Version 1.0 of the specification was released on 17 October 2008
  - <http://www.openarchives.org/ore/1.0/>
- The goal of OAI-ORE is to
  - expose the rich content in aggregations
- to applications supporting
  - authoring, deposit, exchange, visualization, reuse, and preservation

# OAI-ORE



- Specification is created around the Object Reuse and Exchange Model which introduces the Resource Map (ReM)
- ReM associates an identity with aggregations of resources
- Aggregations (compound information objects) may combine distributed resources with multiple media types

# OAI-ORE



- Aggregations (compound information objects) may combine distributed resources with multiple media types
- Aggregation can be a part of other aggregations

# OAI: Object Reuse and Exchange



1. Browser address bar: <http://arxiv.org/abs/astro-ph/0601007>

2. Download options: PostScript, PDF, Other formats

3. Article title: **Parametrization of K-essence and Its Kinetic Term**

4. Author names: Hui Li, Zong-Kuan Guo, Yuan-Zhong Zhang

5. Submission info: (Submitted on 31 Dec 2005 (v1), last revised 18 Jan 2006 (this version, v2))

6. DOI: 10.1142/S0217732306019475

7. Submission history: [v1] Sat, 31 Dec 2005 04:01:23 GMT (20kb)

8. Current browse context: astro-ph > new | recent | 0601

9. References & Citations: SLAC-SPIRES HEP, NASA ADS, CiteBase

Article text: We construct the non-canonical kinetic term of a k-essence field directly from the effective equation of state function  $w_k(z)$ , which describes the properties of the dark energy. Adopting the usual parametrizations of equation of state we numerically reproduce the shape of the non-canonical kinetic term and discuss some features of the constructed form of k-essence.

Comments: 8 pages, 1 figure; accepted by Mod. Phys. Lett. A; minor changes to references

Subjects: Astrophysics (astro-ph)

Journal reference: Mod.Phys.Lett. A21 (2006) 1683-1690

Cite as: arXiv:astro-ph/0601007v2

Submission history

From: Hui Li [view email]

[v1] Sat, 31 Dec 2005 04:01:23 GMT (20kb)

[v2] Wed, 18 Jan 2006 06:16:15 GMT (20kb)

Which authors of this paper are endorsers?

Link back to: arXiv, form interface, contact.

Source: ORE User Guide – Primer

(<http://www.openarchives.org/ore/1.0/primer.html>)

# OAI-ORE



- Resource Maps may be written in several different formats
  - Atom feed, RDF/XML, RDFa and others
- Example ReM – Atom feed:
  - [http://en.wikipedia.org/wiki/Open\\_Archives\\_Initiative\\_Object\\_Reuse\\_and\\_Exchange#Resource\\_Maps](http://en.wikipedia.org/wiki/Open_Archives_Initiative_Object_Reuse_and_Exchange#Resource_Maps)

# OAI-ORE



- OAI-ORE will co-exist within the OAI-PMH
- ORE is intended to complement the narrower metadata focus of OAI-PMH
- ORE is now studied by different communities its application goes beyond scholarly communication
  - ORE gives a chance to easily migrate whole repositories

# OAI-ORE - Tools



- ORE Atom Resource Map Validator
  - <http://african.lanl.gov/ovalnet/validate.jsp>
- Full list of available tools :
  - <http://www.openarchives.org/ore/1.0/tools.html>

# Crowdsourcing

- Community is very important thing nowadays
- Community management is becoming a separate aspect of any project management
- Digital libraries should also attract community
  - User generated content can enrich resources
  - People can correct mistakes

# Crowdsourcing

- Term “Crowdsourcing” is neologistic compound of Crowd and Outsourcing
- It is the act of taking tasks traditionally performed by an employee or contractor

# Crowdsourcing

## The Crowdsourcing Process *In Eight Steps*



Image by Darin C. Brabham | [www.darincbrabham.com](http://www.darincbrabham.com)

Source: <http://en.wikipedia.org/wiki/Crowdsourcing>

# Crowdsourcing

- How crowdsourcing can be utilized in digital libraries?
  - Flickr : The Commons
    - <http://www.flickr.com/commons>
  - Australian Newspapers Digitization Program
    - <http://www.nla.gov.au/ndp/>
  - Oxford's Great War Poetry Archive
    - <http://www.thegreatwatarchive.org/>

# Australian Newspapers

- Project coordinated by National Library of Australia
- It is intended to give access to Australian newspapers published between 1803 and 1954
- Goal is to give a free access to 40 million of articles
- User will be able to perform a full-text search for all articles

# Australian Newspapers

- They employed different means to cooperate with users, including :
  - comments
  - tags
  - OCR correction
- During the first 12 weeks 1200 people registered in portal and performed some OCR corrections
- This resulted in 700 000 lines corrected in 50 000 articles

# Australian Newspapers

- Users are also submitting different information like:
  - Additional remarks about people/places/situations mentioned in the article
  - Their remarks about scan quality
  - Problems associated with using a portal
  - Information about errors made by other users

# Australian Newspapers

- Why people are investing their time in such a project?
  - “We are sick of doing housework!”
  - “I enjoy typing, want to do something useful and find the content fascinating”
- Is it addictive?
  - Most of participants say – yes, it is
- More user statements at:
  - [http://www.nla.gov.au/ndp/news\\_and\\_events/documents/NDP\\_IMPACT\\_MANYHANDS\\_April2009.ppt](http://www.nla.gov.au/ndp/news_and_events/documents/NDP_IMPACT_MANYHANDS_April2009.ppt)

## FIND AN ARTICLE

[Advanced Search](#)

**Search Articles**

## FIND AN ISSUE

### by Title

1. The Argus
2. The Courier-Mail
3. The Mercury
4. The Perth Gazette and ...
5. The South Australian ...

**Show all titles**

### by State

NT  QLD  
WA NSW  
SA ACT  
TAS VIC

### by Date

1803

JAN	FEB	MAR	APR
MAY	JUN	JUL	AUG
SEP	OCT	NOV	DEC

S	M	T	W	T	F	S

## ON THIS DAY

ARGUS (MELBOURNE, VIC.), WEDNESDAY 11 MARCH 1931

**Navigation tips** for the example newspaper page below:

**Scroll** with the scrollbars or your scrollwheel.

**Pan** by clicking and dragging the image.

**Zoom** with the zoom controls in the bottom right.

[Read this article](#)

an important part in the main-  
ality. Therefore every woman  
interested in a special article  
is week in "The Australasian"  
ages, which describes the latest

**PATTERN SERVICE.**



**Sailing on,  
and on-and on-**

**JUST as the Ship** **ZOOM**   
Leading the sea

## USER LOGIN

username

password

**Login**

## TOP TEXT-CORRECTORS

1. jhempenstall (130077)
2. cmdevine (111734)
3. fwalker13 (109950)
4. maurielyn (93640)
5. John.F.Hall (84616)

## RECENT COMMENTS

The final paragraph on Page 2 (staring "...  
created 2009-03-11 12:44:38.0 by AuFOL

Misses the GURNEY birth entry

created 2009-03-11 10:58:10.0 by sparrowmickey

## RECENT TAGS

A B Tress Ambrose Campbell Carmichael >> All

G W Fuller George Lawrence Fuller tags

George Skelton Yuill

Print Save as PDF Save as Image

Cite: http://nla.gov.au/nla.news-article3268358

Tags (Keywords) Add/Edit Tags
titanic sinking

Comments Add New Comment
Hide Comments

(Edit) This is the first report of the sinking of the Titanic in an Australian paper. The passengers were not all taken off safely as is reported, most drowned. Further reports in the paper give the correct details.
- created 2009-03-10 15:23:33.0 by rholley

ELECTRONICALLY TRANSLATED TEXT
Why may this text have mistakes? Help fix this text!

Text [corrected most recently by cmdevine - Show corrections]

NEWS & NOTES.

Miss Clapp, who for over two years has been on the nursing staff of the Darwin Hospital, departed south by the "Empire".

Among the forty passengers by the "Mataram" there arrived Mr. T. E. Day, Chief Surveyor and nine men. for the Survey Department.

The Darwin District Council held its fortnightly meeting on Tuesday evening last, but owing to poor attendance no important business was dealt with.

WRECK OF THE "TITANIC."
April 16.

THE Atlantic liner "Titanic," the new floating city of 45,000 tons, making her maiden voyage from Southampton to New York, struck an iceberg to the south of Newfoundland. There were 2000 people on board. The weather was calm and all the passengers were taken off. Assistance was summoned by wireless telegrams, and the Titanic was reported to be sinking. Several liners hastened to her aid:

SHIP FOUNDERED.
April 16:

THE ship "Songvaar" foundered at Port Victoria, South Australia, with a valuable cargo of wheat. The cause is unknown.

FEDERAL SURPLUS.
April 16.

THE Federal Government expects to...



View entire page

ZOOM [zoom controls]

## Text corrections

### [General Orders.](#)

The Sydney Gazette and New South Wales Advertiser Saturday 5 March 1803, page 1

Changed	By	Old Lines	New Lines
<a href="#">user:public:Bdamokos</a>		detachments and labouiiing people at Castle	detachments and labouring people at Castle-
<a href="#">user:public:Bdamokos</a>		jrelieves him ; the said Orders are also to be	relieves him ; the said Orders are also to be
<a href="#">user:public:Bdamokos</a>		-m%, ked off in the Extracts he is furnished	marked off in the Extracts he is furnished
<a href="#">user:wcathro</a>		Settlers at Hawkeibury, fiom the vexatious	Settlers at Hawkesbury, from the vexatious
<a href="#">user:wcathro</a>		General Order	General Orders
<a href="#">user:wcathro</a>		rs<<	General Order
anonymous		Boat receives more grain than the vessesl	Boat receives more grain than the vessel
<a href="#">user:lcho</a>		m E PE AT ED Complaints hiving been made of the great loíies fuílained by the	REPEATED Complaints hiving been made of the great loíies fuílained by the

# Other examples

- Virtual manuscript room
  - <http://vmr.bham.ac.uk>
- Such a online virtual laboratory is the only chance to :
  - analyze those manuscripts
  - prepare transcription



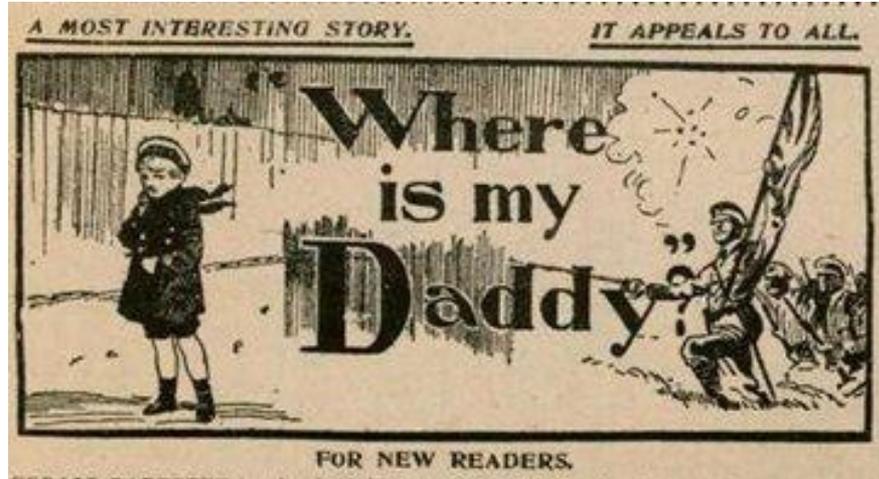
# Oxford's Great War Poetry Archive

- Project coordinated by University of Oxford
  - <http://www.thegreatwararchive.org/>
- Last only 3 months from 8.03.2008 till 11.11.2008.
- Goal:
  - Create valuable collection at low cost
  - Avoid institutionalized digitization

# Oxford's Great War Poetry Archive

- Community Collection Building
- Everyone could propose resource which might be added to collection:
  - Not only poetry but also letters, old pictures and stories associated with them
- Project was aimed to facilitate creation of educational resources like
  - Podcasts, video, articles etc.
- It also gathered information about existing educational resources

# Oxford's Great War Poetry Archive



# Oxford's Great War Poetry Archive

- Volunteer digitization
  - They created a group at Flickr.com where people were able to add their content associated with a topic:
    - *"I have recently inherited my (german) grandfather's old photo album from WW1. I have posted some of these photos onto flickr already but not to any groups [...] Are you interested?"*
      - <http://www.flickr.com/groups/greatwararchive/discuss/72157605915465052/>
- During 3 month they gathered 6 500 of objects
- This group is still open for submission (till now they gathered 2 000 additional pictures)

# Oxford's Great War Poetry Archive

---

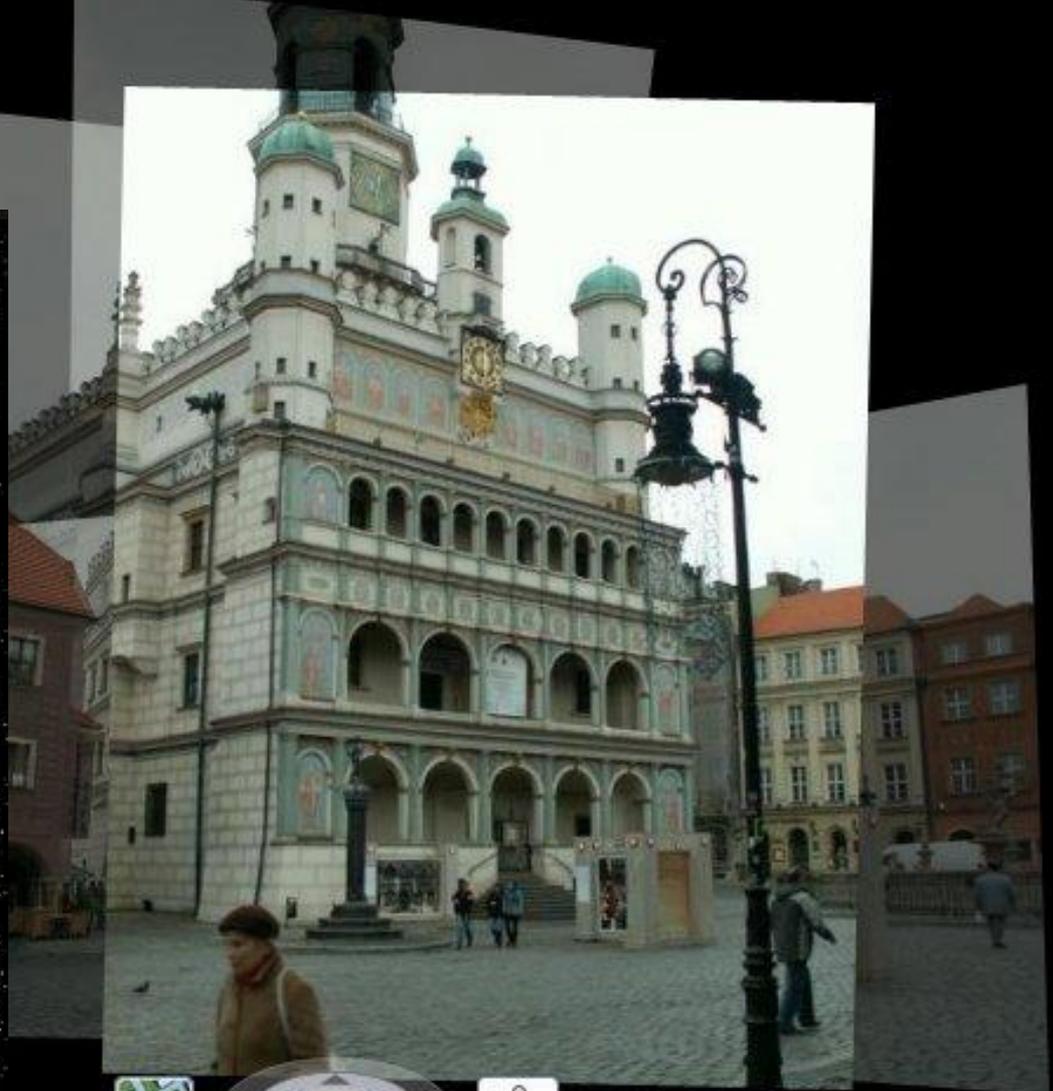
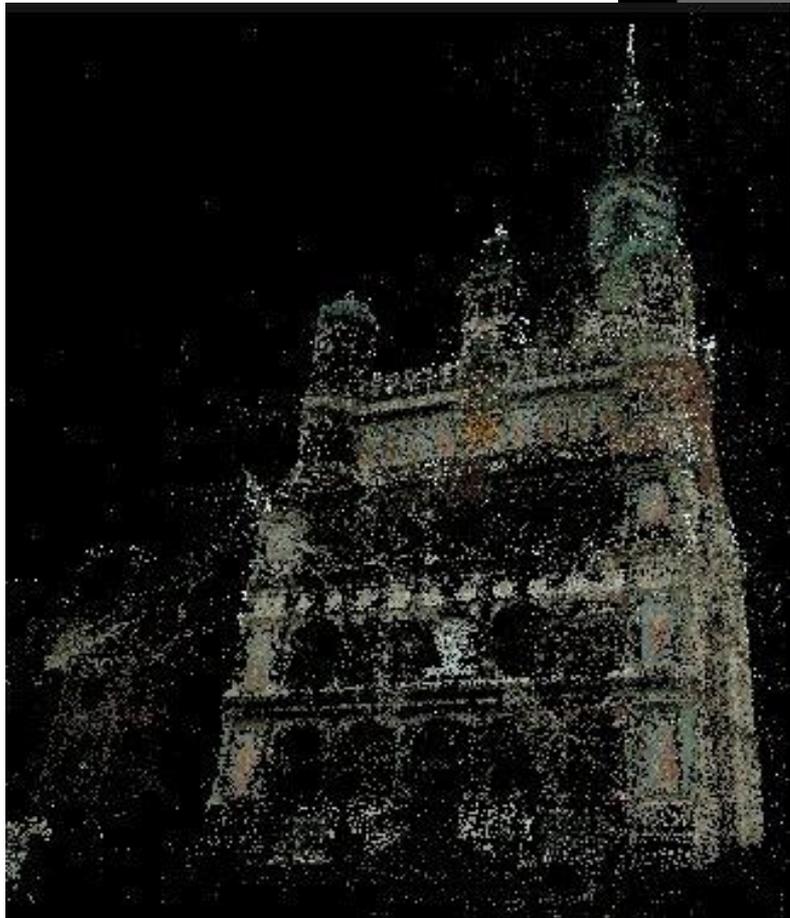
- This approach reduced the cost of digitization even 10 times

# Conclusions

- Things are changing very quickly
- New technologies, media are appearing all the time
- Digital Librarians have a special role in this process
- New technologies creates new possibilities and challenges

Tartu. Estonian Historical Museum  
collection: EAM N5635:52  
Modern rephotography by Vahur Puik, 2009





# Conclusions

- The question is which of them are the most important, which should be preserved
- Community participation is a great chance but engaging internet users is sometimes very hard task

# Q&A

---