Tony Kent Strix Annual Lecture - 20 October 2017

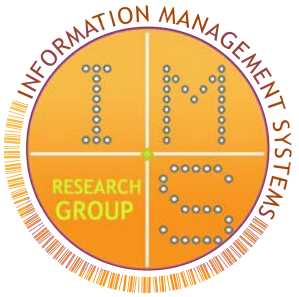
Behind the Scenes of Research and Innovation

Maristella Agosti

Information Management Systems Research Group (IMS)

Department of Information Engineering

University of Padua, Italy



From Research to Innovation

- DUO: An Innovative OPAC (online public access catalogue for libraries)
- FAST: Bringing Annotations into Digital Libraries
- DIRECT: IR Experimental Data Management

DUO:
An Innovative OPAC

The Italian Library Automation Project and the OPAC

- The Italian national project of library automation, called SBN - *Servizio Bibliotecario Nazionale*, is an advanced library automation project started in 1970s
  - Different library automation systems at national/regional/local level cooperating in a networked/hierarchical organisation
- Until late in the 1980s
  - The public online access to bibliographic data was not available, only traditional card catalogues were in use

OPAC Access at the University of Padua

- The University of Padua became a node of the SBN project in the late 1980s
- At that time, there was much interest in OPAC
  - A first indication that information retrieval might start to interest the general public of libraries
- We launched a project for a third generation OPAC with advanced library catalogue and IR functions

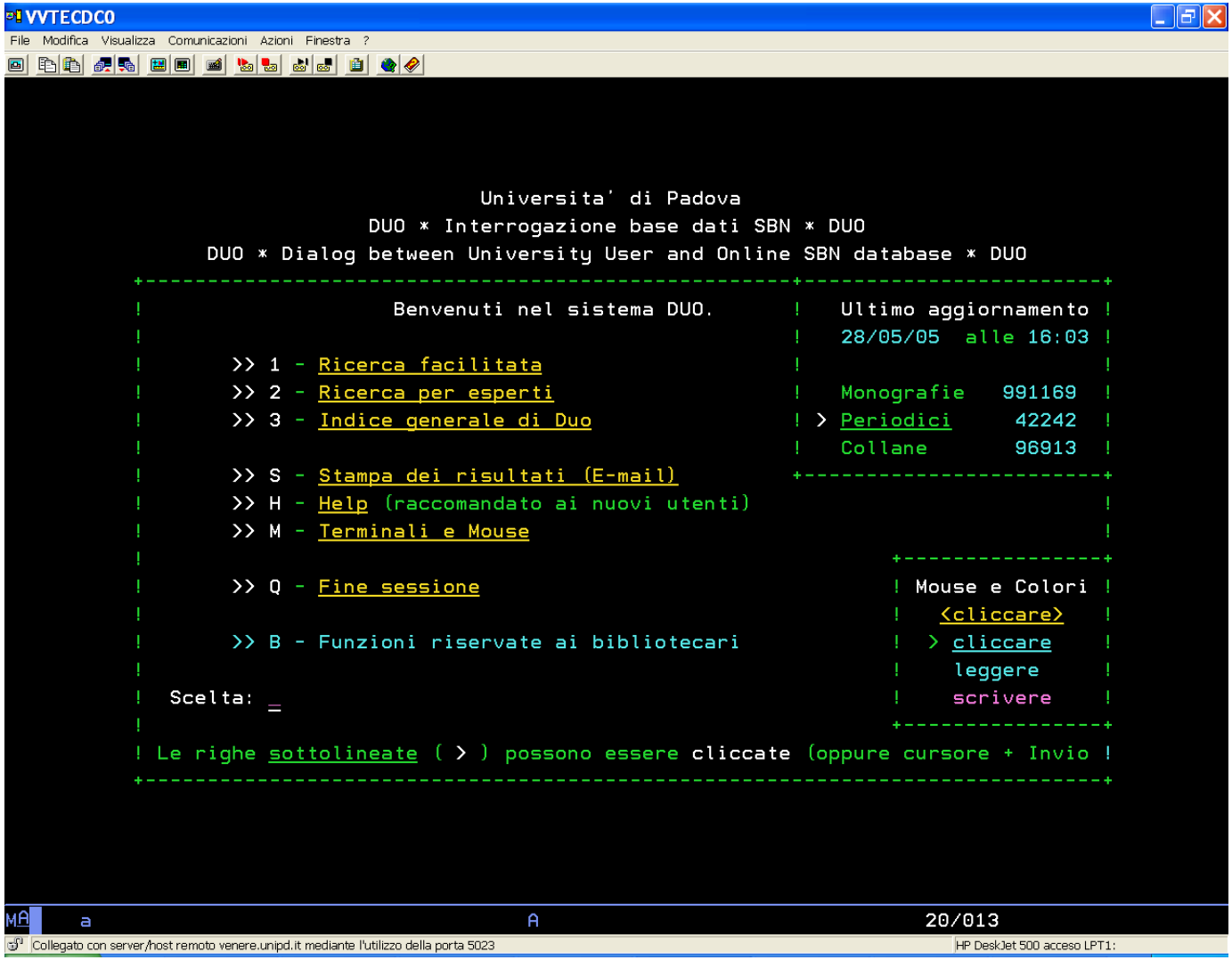
DUO: The OPAC of the University of Padua

- Innovative search functionalities
  - multi-fielded search
  - taxonomies/faceted search
  - fully unstructured document search

  over a co-operative multi-discipline library catalogue database

- Prototype available to users in June 1991
- DUO was openly available on the Internet through the “OPAC” public login using Telnet
- Possibly the first OPAC openly accessible on the Internet free of charge - the Web did not exist at that time

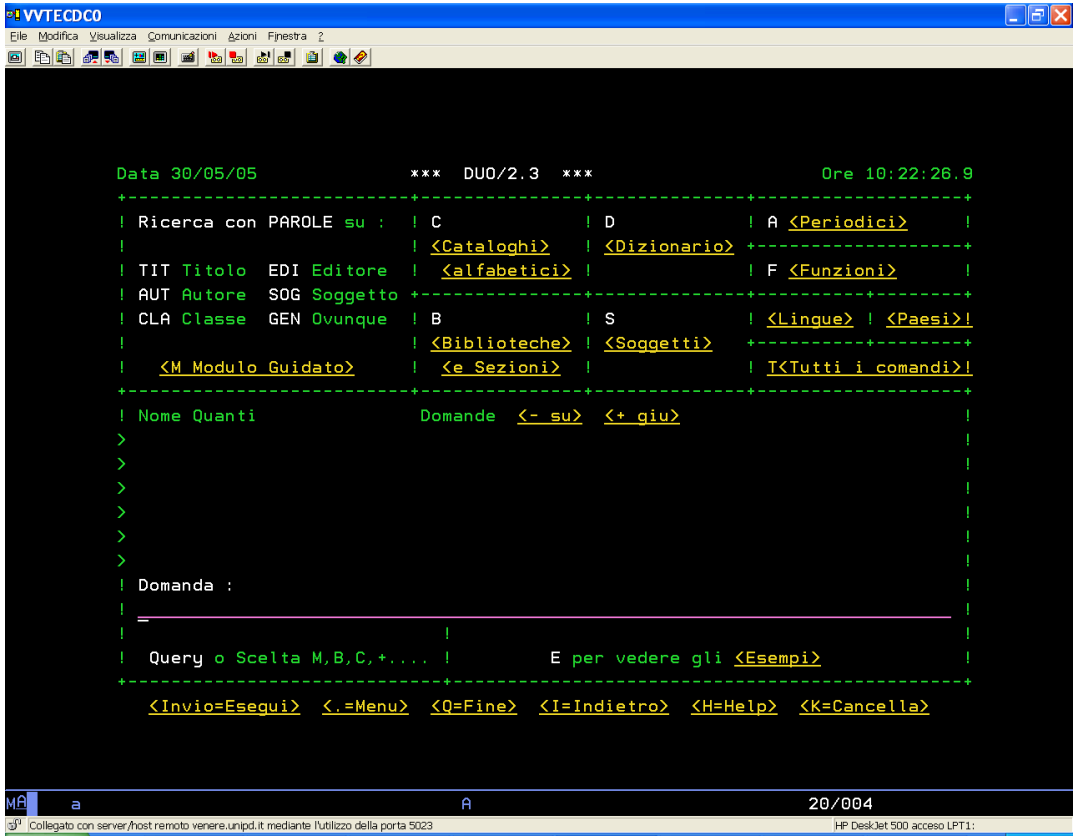
The OPAC DUO Interface (in Italian)



Maristella Agosti

A Text Box for Query Input

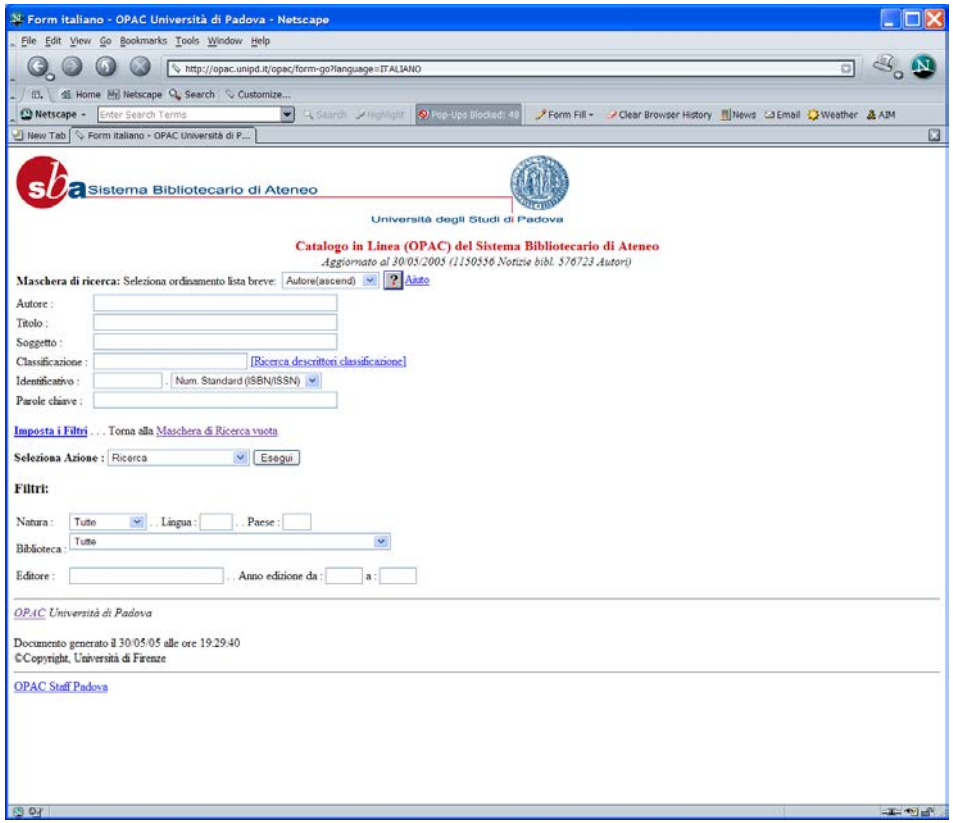
The text box for free query input was innovative



We studied Okapi – probably the first system with a text box for free query input

Maristella Agosti

“Evolution” of DUO: Access to the Catalogue through the Web



- The time was not ripe for Web applications: the IR functions were lost

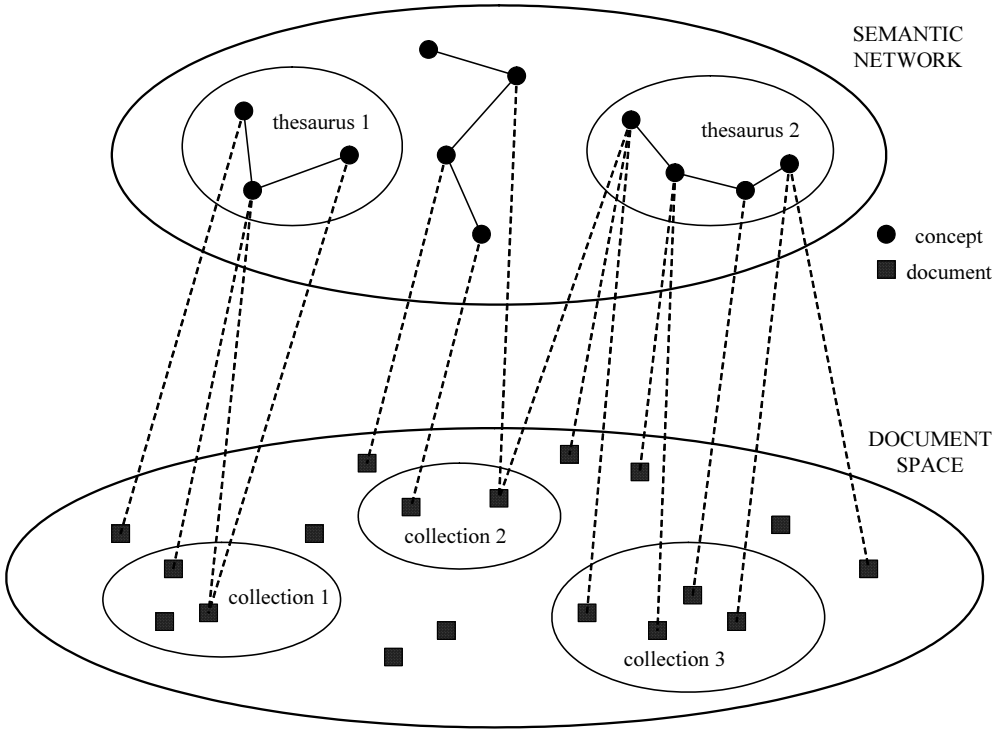
The Birth of the Digital Library Area

- The Library Automation community realises the lack of computer science and engineering knowledge
- The area of Digital Library starts in those years as a new scientific area
  - In USA - Digital Libraries Initiatives (DLI-1 and DLI-2) of the National Science Foundation (from late 1993)
  - In Europe - A group of projects supported by the European Commission under the 4th, 5th and 6th Framework Programme named DELOS Working Group 1996-99, first DELOS Network of Excellence 2000-2003, and DELOS Network of Excellence for Digital Libraries 2004-2007
- It is an area of confluence: library automation, database management, information retrieval, the Web, …

FAST:
Bringing Annotations into Digital Libraries

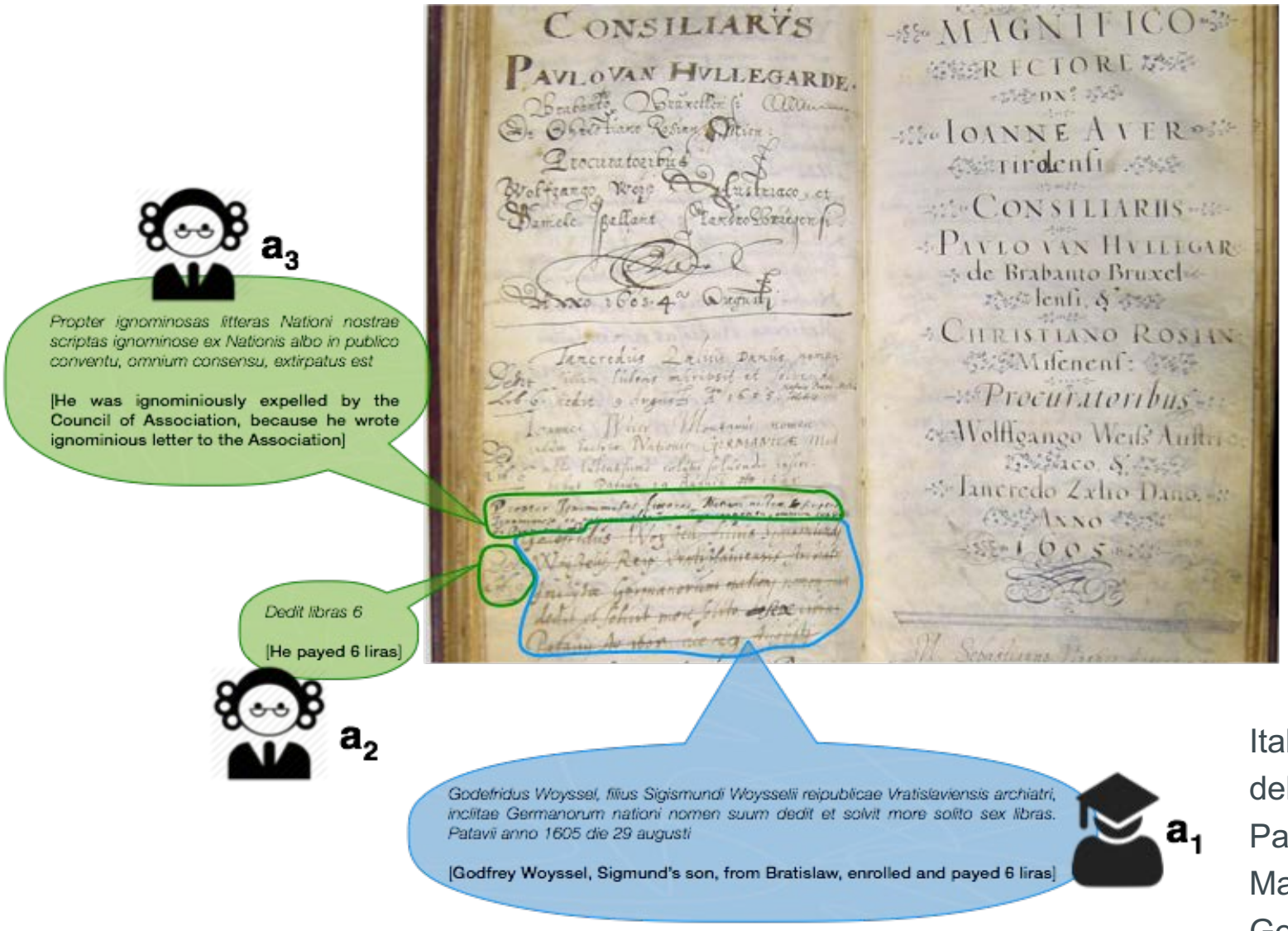
Background Research Experience: Hypertext Information Retrieval – 1980/1990

- The EXPLICIT Model for Hypertext IR



Maristella Agosti

Historical Annotations: Padua University

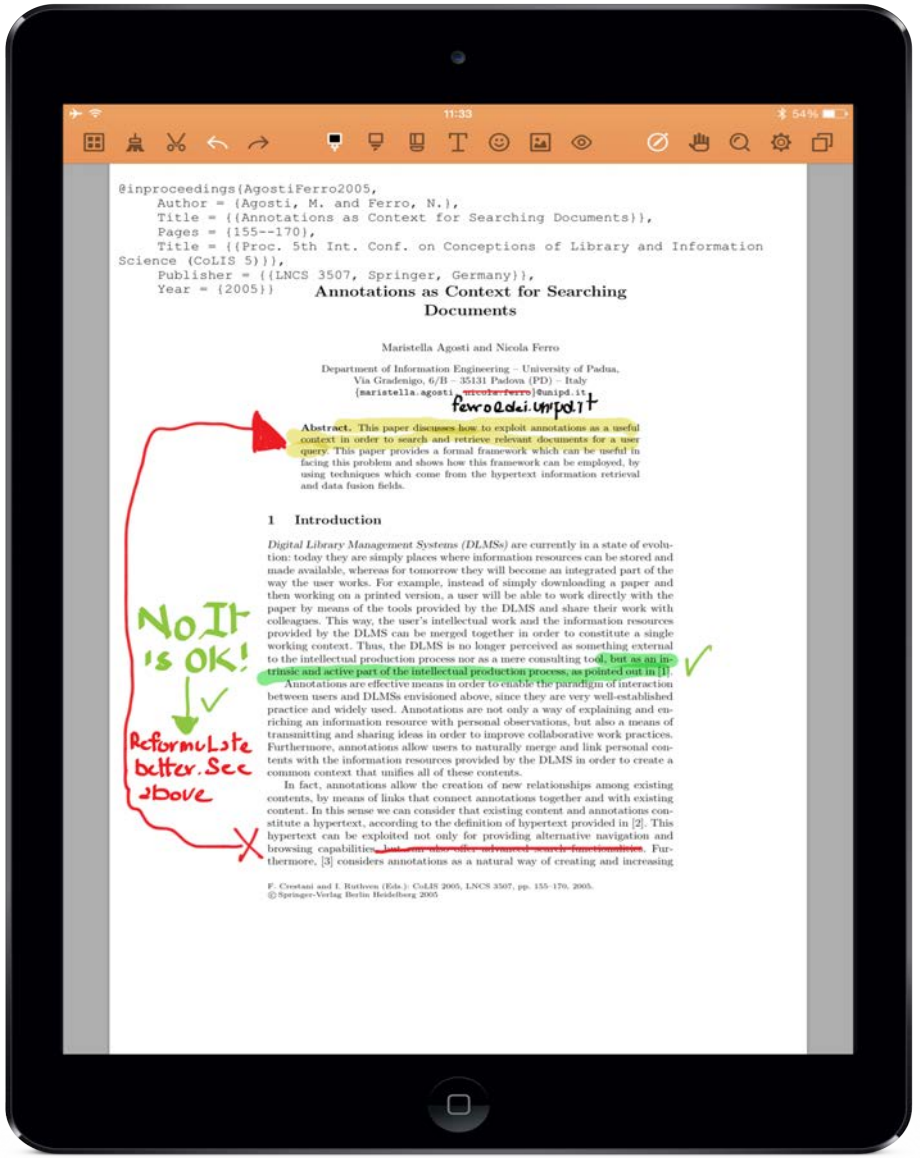


Italia, Padova, Archivio dell'Università di Padova, Archivio antico, Matricula Nationis Germanicae artistarum, reg. 465, c. 69v

Maristella Agosti

Key Issues - also on Today Tablets

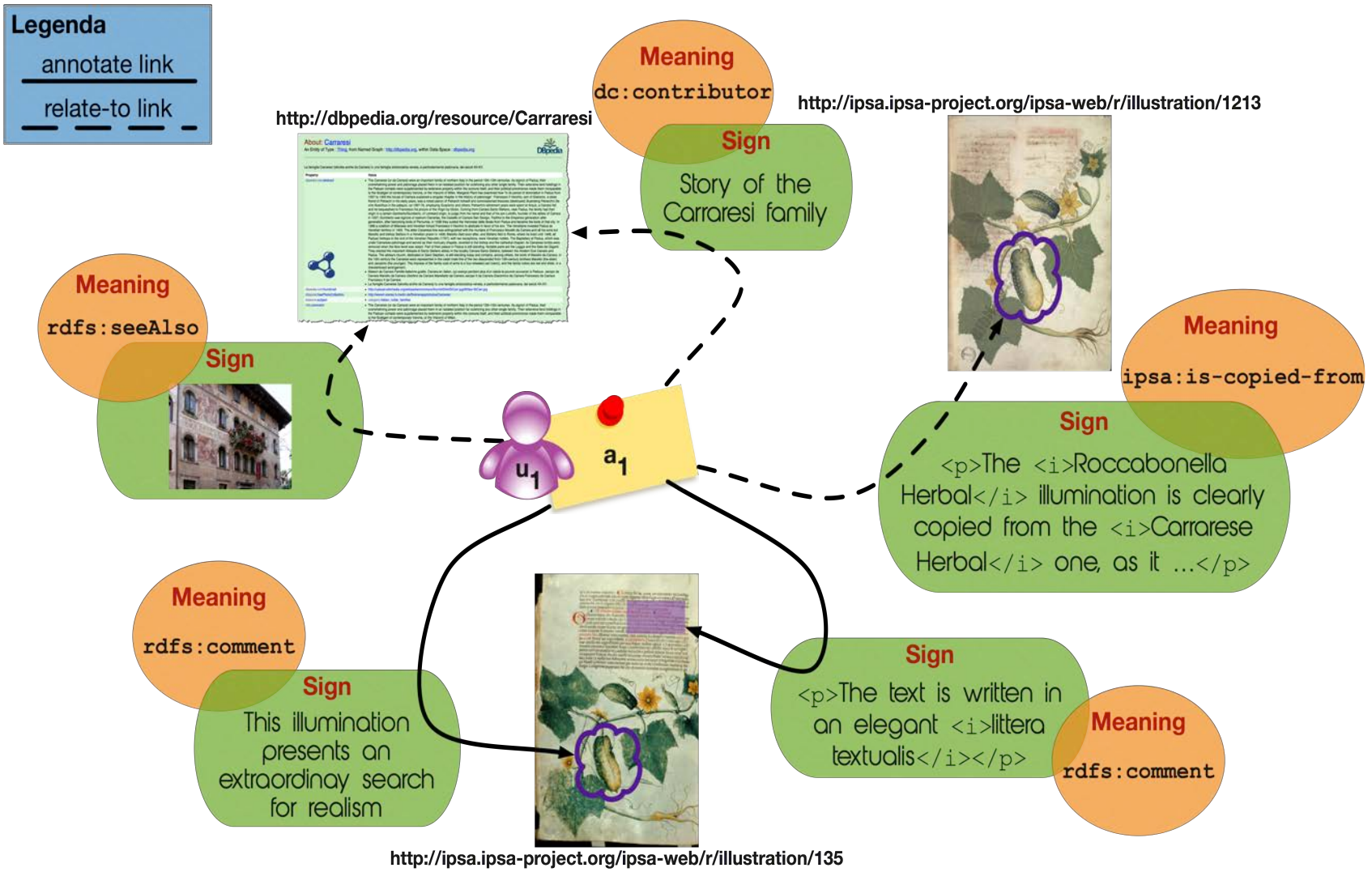
- Annotations are embedded in the annotated document
- Annotations semantics is not explicit or hard-coded
- Annotations are not related one to the other
- All the annotations have the same scope
- Annotations are not searchable



What to Expect from Annotations?

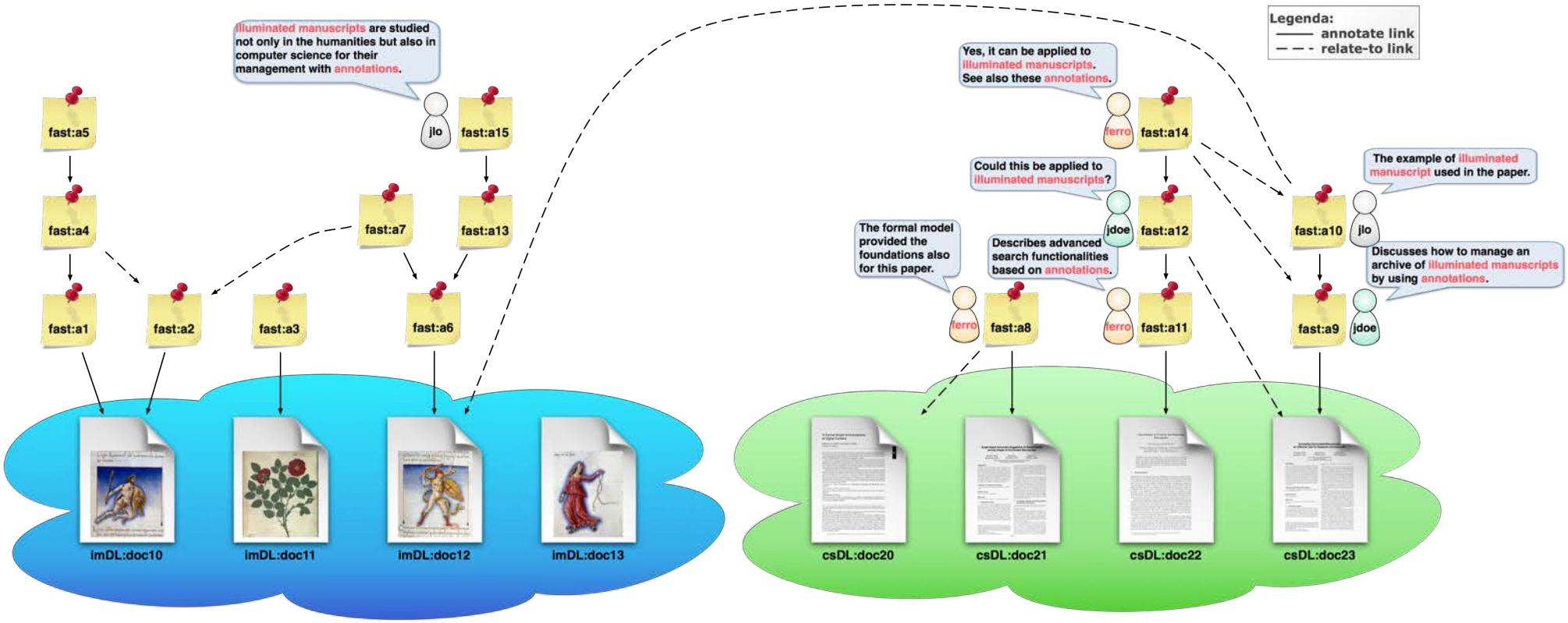
- A collaborative tool for user generated content
- Open, distributed and interoperable among different systems (the Web, digital libraries, digital archives, …)
- Able to engage research communities, foster their research work and transfer knowledge to students and the general public

Annotation Model



Maristella Agosti

The Document-Annotation Hypertext

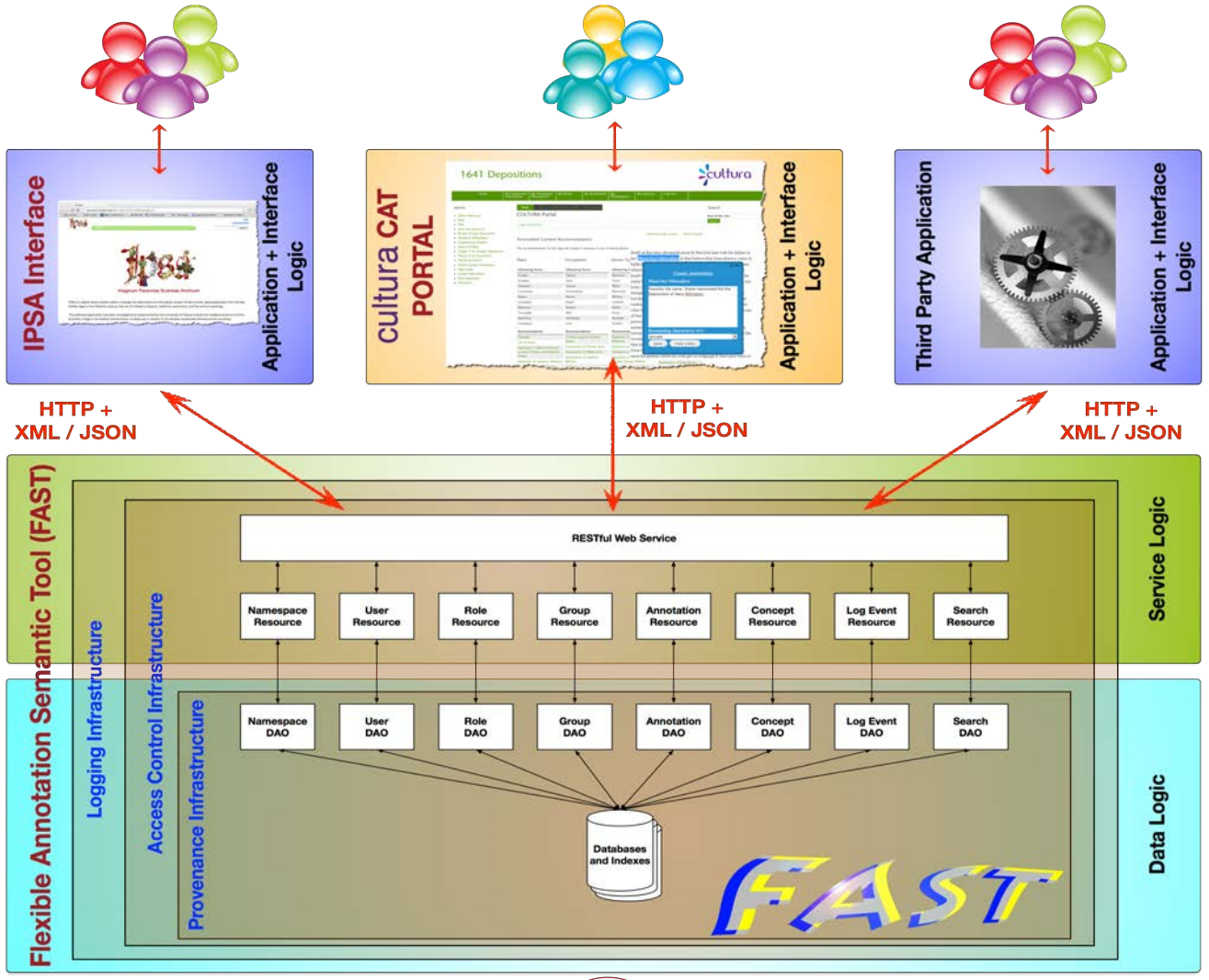


imDL: Illuminated Manuscripts Digital Library

csDL: Computer Science Digital Library

- Search by using annotations: exact match, best match, and navigation of the hypertext

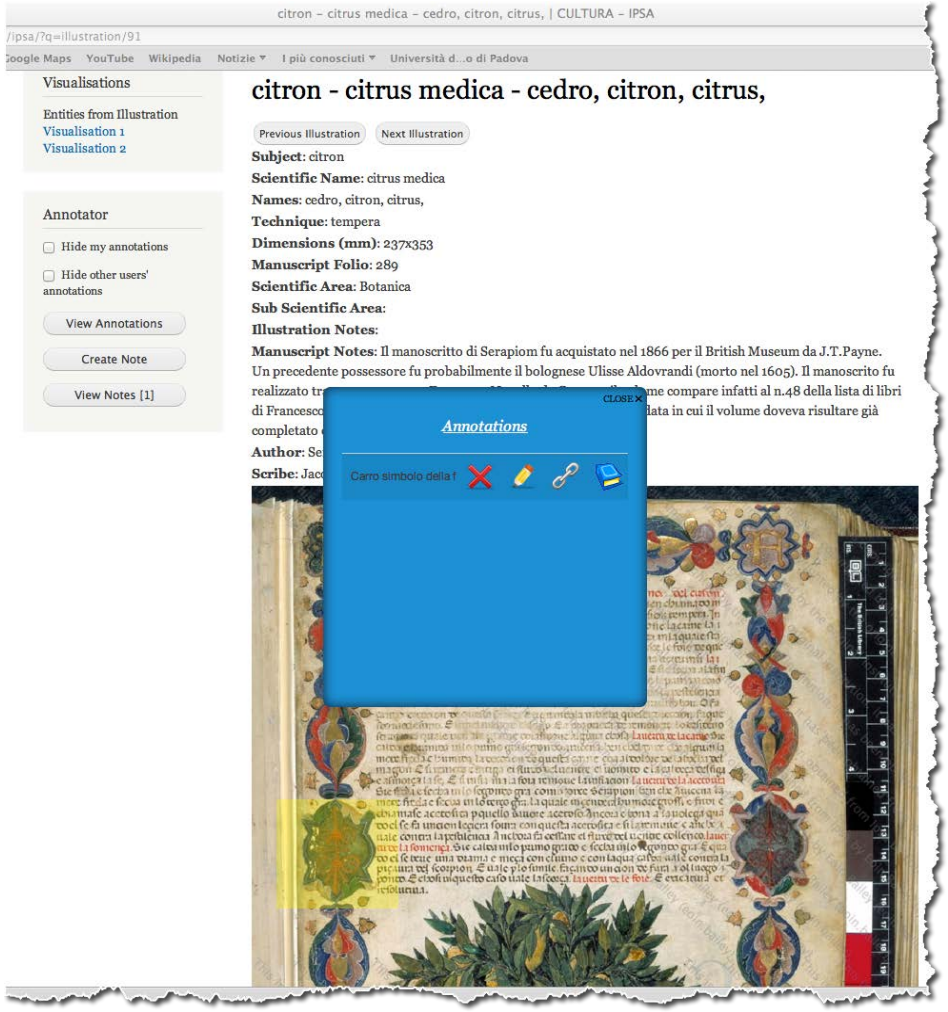
FAST (Flexible Annotation Semantic Tool): a Tool to Innovate



Maristella Agosti

An Example of Transfer of Innovation: Annotations in the CULTURA Project

- The CULTURA project
  - innovative environment for users with a range of different expertise
  - users can collaboratively explore, interrogate and interpret complex and diverse digital cultural heritage collections
- Use cases
  - IPSA: a digital archive of illuminated manuscripts produced in northern Italy during the 14th and 15th centuries
  - The 1641 Depositions: the documents contain witness testimonies from men and women from all over Ireland and report on the rebellion of October 1641



Maristella Agosti

Considerations on the Annotations Effort

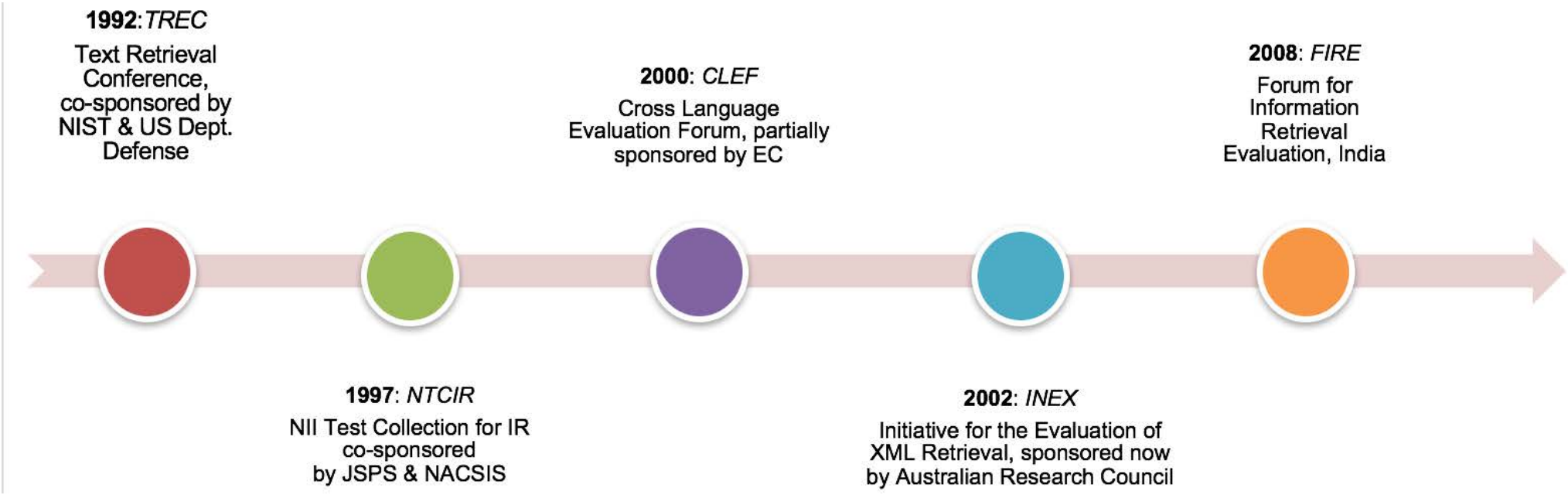
- Modelling, managing and searching annotations is a challenging research problem
  - 5 years to achieve a comprehensive formal model
  - 2 more years to achieve search over/by annotations
  - impact on the field - see W3C Open Annotation Collaboration - OAC, and only for Web annotations
- Developing a fully fledged annotation service is a demanding activity
  - 7 years to develop the FAST service and integrate it into several digital library systems in effective use

DIRECT:
IR Experimental Data Management

“Traditional” IR Evaluation

- IR is intrinsically probabilistic and not deterministic, so the evaluation of effectiveness is necessary (to my knowledge, the first area of computer science and engineering where effectiveness evaluation was conducted)
- IR evaluation is based on a comparative evaluation approach in which system performances are compared according to the Cranfield methodology, which makes use of test collections: $C = \{ D, T, RJ \}$
- A test collection C allows the comparison of information access systems according to measurements which quantify their performances
- Main goals of a test collection
  - to provide a common test-bed to be indexed and searched by information access systems
  - to guarantee the possibility of replicating the experiments

Large-Scale Evaluation Initiatives



- Evaluation initiatives have been relying mainly on the traditional Cranfield methodology, focusing on:
  - the creation of comparable experiments
  - the evaluation of performance

What is Missing in the Cranfield Paradigm?

- The “Cranfield” evaluation initiatives produce different kinds of valuable experimental data, but …

- Scientific data should be properly managed and tracked

- Scientific data should be curated and progressively enriched by adding further analyses and interpretations on them

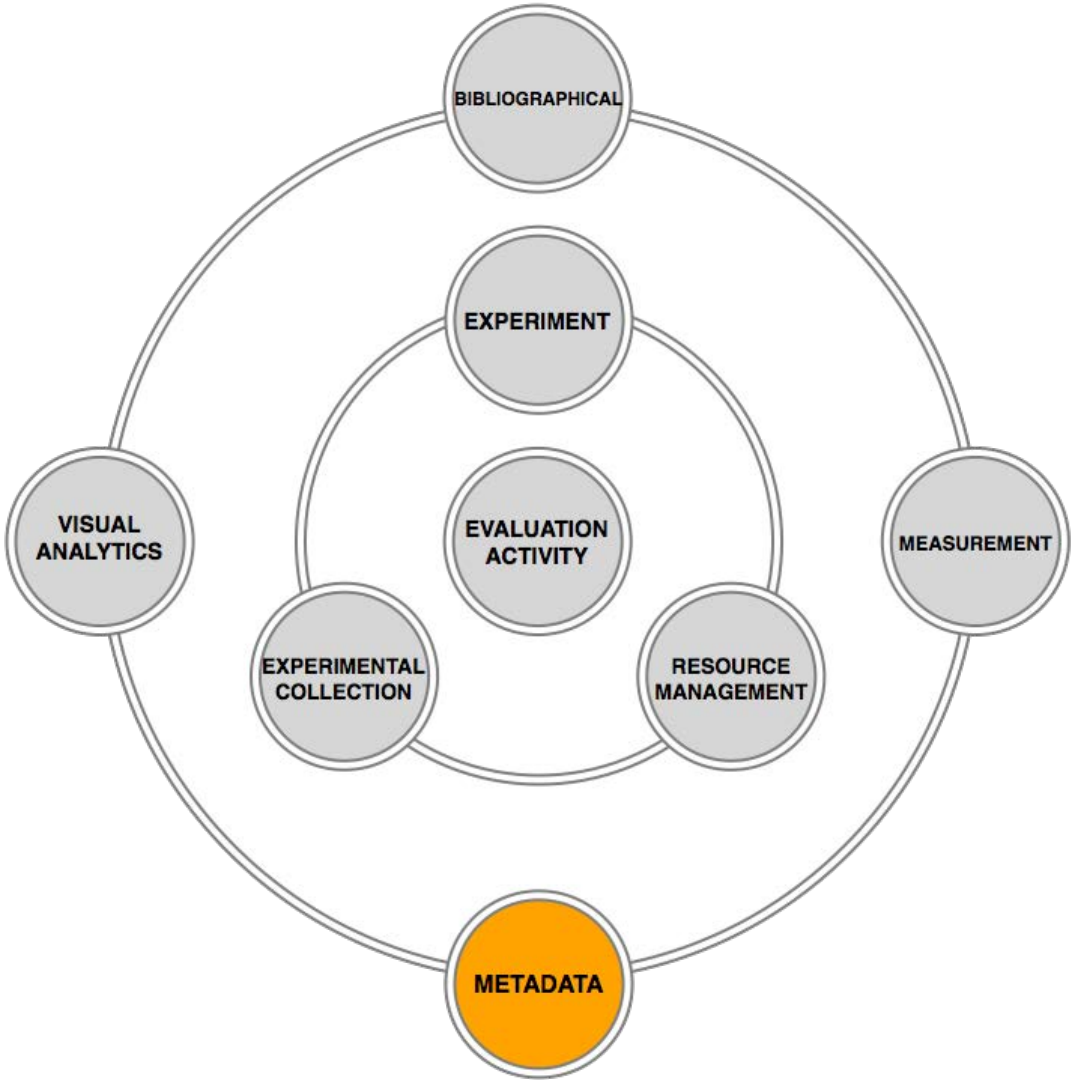
Extensions to the Cranfield Paradigm for Scientific Data Management

- Modelling and managing the valuable scientific data produced during an evaluation campaign
- Citing data to make IR experimental data “first class citizens”
- Improving cooperation and facilitating the transfer of scientific and innovative results from research groups to the industrial sector

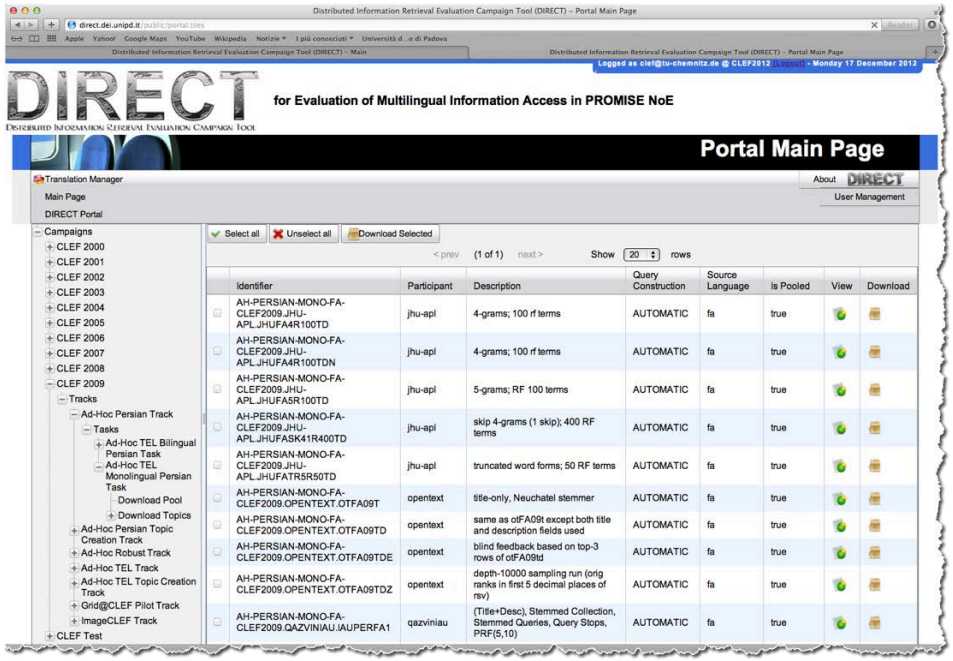
The DIRECT Approach

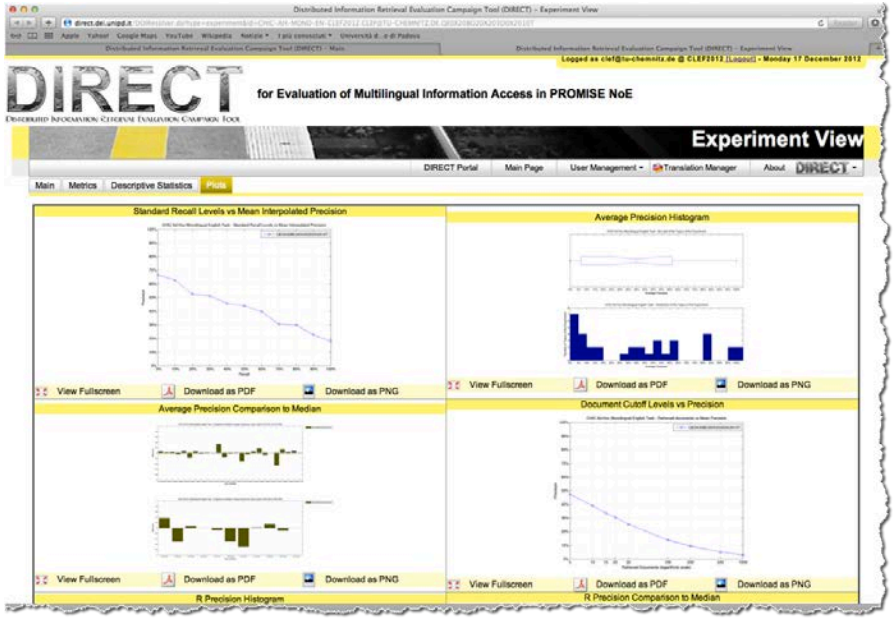
- Introduce a **conceptual model**
- Develop **common metadata formats**
- Adopt a **unique identification mechanism**
- Provide **common tools for statistical analyses**
- Provide a **Digital Library System** (DLS) to manage IR scientific data named DIRECT (Distributed Information Retrieval Evaluation Campaign Tool)
- Give organizations responsible for evaluation initiatives an active role in the process

The DIRECT Approach: Modelling Areas



The DIRECT Web Application





Remarks on Advanced IR Evaluation

- To do research and innovation in IR a diversified knowledge is needed in other disciplinary sectors, including just to name a few: database management, digital libraries, statistics, probability, information science, …
- Both the academic community and the private sector should work towards and foster the transparency of scientific results to ensure their reproducibility

Thank you for your attention

Questions?

References on OPAC and DUO

- M. Agosti, M. Masotti, A.M. Moressa. An Online Public Access Catalogue (OPAC) for University Library End-users Using TRS: project and prototype. *Proc. Software AG's European Users' Conference*, Hamburg, Germany, 1990, Vol.1, Paper N.52

- M. Agosti, M. Masotti. Design of an OPAC database to permit different subject searching accesses in a multi-disciplines universities library catalogue database. In: N. J. Belkin, P. Ingwersen, A. M. Pejtersen (Eds.). *Proc. of the 15th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*. Copenhagen, Denmark, ACM, 1992, 245-255

- S. Robertson, On the history of evaluation in IR. *Journal of Information Science,* 2008, 34(4), 439-456

- S. Walker. Improving subject access painlessly: recent work on the Okapi online catalogue projects. *Program*, 1988, 22(1), 21-31

References on Hypertext IR

- M. Agosti, R. Colotti, G. Gradenigo. A two-level hypertext retrieval model for legal data. In: A Bookstein, Y. Chiaramella, G. Salton, V. V. Raghavan (Eds). *Proc. ACM SIGIR 1991*, Chicago, USA, 316-325

- M. Agosti, G. Gradenigo, P.G. Marchetti. A Hypertext Environment for Interacting with Large Textual Databases. *Information Processing & Management*, 1992, 28(3), 371-387

- M. Agosti. Editorial - Hypertext and Information Retrieval. *Information Processing & Management*, 1993, 29(3), 283-285

- M. Agosti, A. Smeaton (Eds). *Information Retrieval and Hypertext*. Kluwer Academic Publishers, Boston, 1996

References on Digital Annotations

- M. Agosti, N. Ferro. A Formal Model of Annotations of Digital Content. *ACM Transactions on Information Systems (TOIS)*, 2008, 26(1):3:1-3:57

- M. Agosti, G. Bonfiglio-Dosio, N. Ferro. A Historical and Contemporary Study on Annotations to Derive Key Features for Systems Design. *International Journal on Digital Libraries*, 2007, 8(1):1-19

- M. Agosti, N. Ferro. Annotations as Context for Searching Documents. In: F. Crestani, I. Ruthven (Eds). *Proc. of* CoLIS 5, LNCS 3507, Springer, Heidelberg, Germany, 2005, 155-170

- N. Ferro. Annotation Search: The FAST Way. In: M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas (Eds). *Proc. 13th ECDL,* LNCS 5714, Springer, Heidelberg, Germany, 2009, 15-26

- M. Agosti, O. Conlan, N. Ferro, C. Hampson, G. Munnelly, G. Interacting with Digital Cultural Heritage Collections via Annotations: The CULTURA Approach. In S. Marinai, K. Marriot (Eds). *Proc. 13th ACM DocEng*, ACM Press, New York, USA, 2013, 13-22

References on the DIRECT Approach

- M. Agosti, G.M. Di Nunzio, N. Ferro. Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In: A. Nardi et al (Eds). *Working Notes for the CLEF 2006 Workshop*, Alicante, Spain, 2006, 11-20

- M. Agosti, G.M. Di Nunzio, N. Ferro. The Importance of Scientific Data Curation for Evaluation Campaigns. In: C. Thanos et al (Eds). *Digital Libraries: Research and Development, First Int. DELOS Conference*, Revised Selected Papers. Springer, Berlin/Heidelberg, Germany, LNCS 4877, 2007, 157-166

- M. Agosti, N. Ferro. Towards an Evaluation Infrastructure for DL Performance Evaluation. In: G. Tsakonas, C. Papatheodorou, C. (Eds). *Evaluation of Digital Libraries: An Insight to Useful Applications and Methods*. Chandos Publishing, Oxford, UK, 2009, 93-120

- M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, G. Silvello. DIRECTions: Design and Specification of an IR Evaluation Infrastructure. In: T. Catarci et al (Eds). *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics - 3rd Int. Conf. of the CLEF Initiative*, CLEF 2012. LNCS 7488, Springer, Berlin Heidelberg, 2012, 88-99