PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

Evaluation Infrastructures

Nicola Ferro

Information Management Systems (IMS) Research Group
Department of Information Engineering (DEI)
University of Padua, Italy

PROMISE Winter School 2013 on Bridging between Information Retrieval and Databases
Bressanone, Italy, 7 February 2013

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191 http://www.promise-noe.eu/

Outline

- Motivations for an evaluation infrastructure
- Modelling
- Architecture and outcomes
- Final remarks

Why an Evaluation Infrastructure?

What is an Infrastructure?

the basic physical and organizational structures and facilities (e.g., buildings, roads, and power supplies) needed for the operation of a society or enterprise

What is an Infrastructure?

the basic physical and organizational structures and facilities (e.g., buildings, roads, and power supplies) needed for the operation of a society or enterprise

What is an Infrastructure?

the basic physical and organizational structures and facilities (e.g., buildings, roads, and power supplies) needed for the operation of a society or enterprise

What is an Infrastructure?

the basic physical and organizational structures and facilities (e.g., buildings, roads, and power supplies) needed for the operation of a society or enterprise

What is an Infrastructure?

the basic physical and organizational structures and facilities (e.g., buildings, roads, and power supplies) needed for the operation of a society or enterprise

What is an Infrastructure?

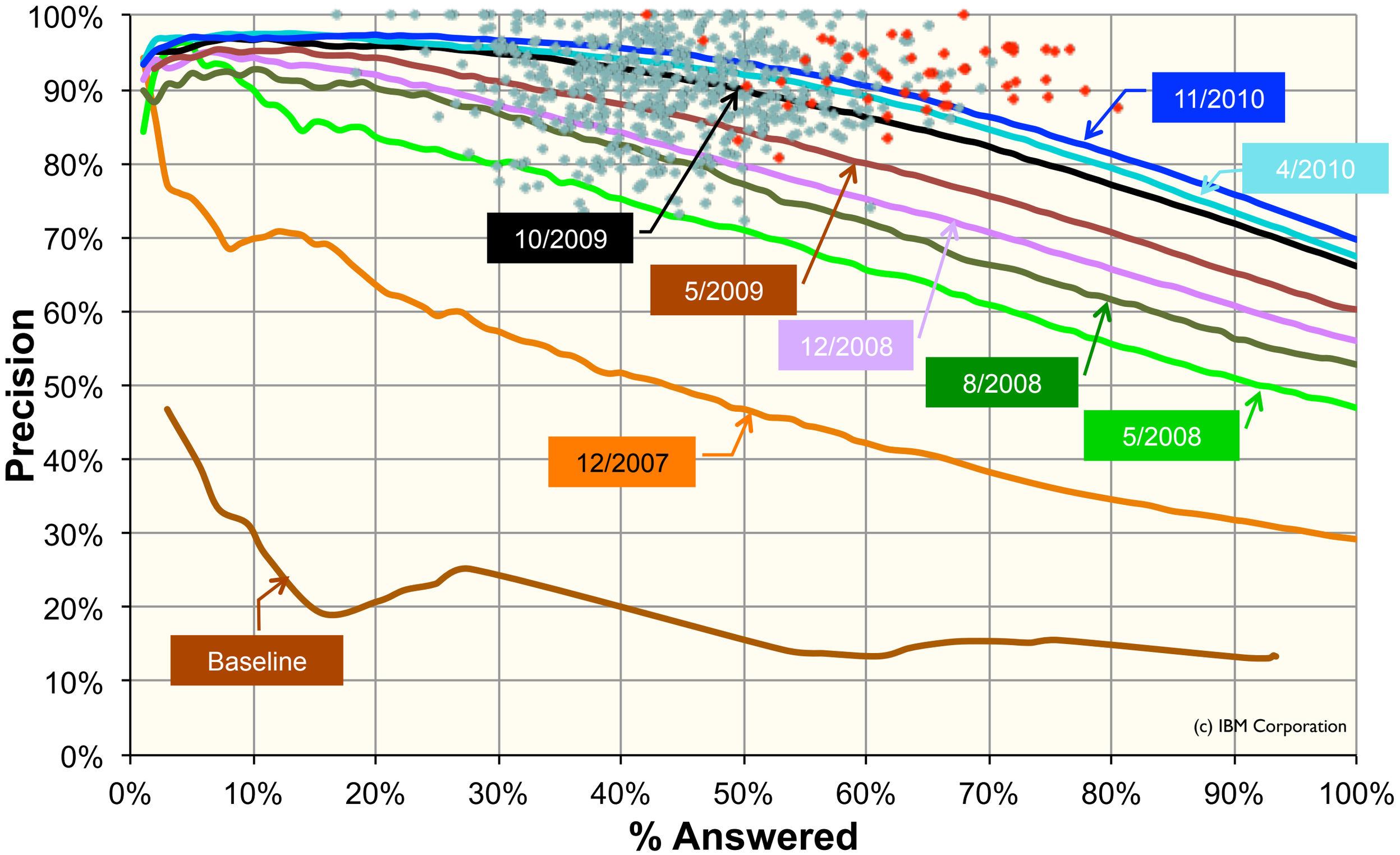
the basic physical and organizational structures and facilities (e.g., buildings, roads, and power supplies) needed for the operation of a society or enterprise

In our case

Why Evaluation?

IBM Watson:
Deep QA Project

Why Evaluation?



What about Society?

- Evaluation initiatives
- Researchers
- Developers and companies
- Stakeholders and adopters

Evaluation Initiatives

Visitor, Participant and Organizer

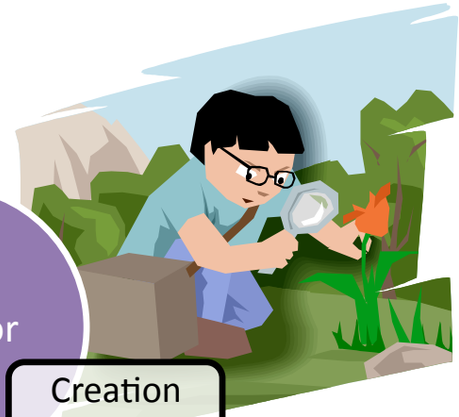
Scientific Production

Organizer

Preparation of Documents

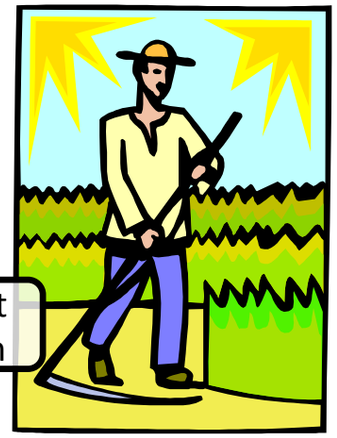
Assessor

Creation of Topics



Participant

Experiment Submission



Organizer

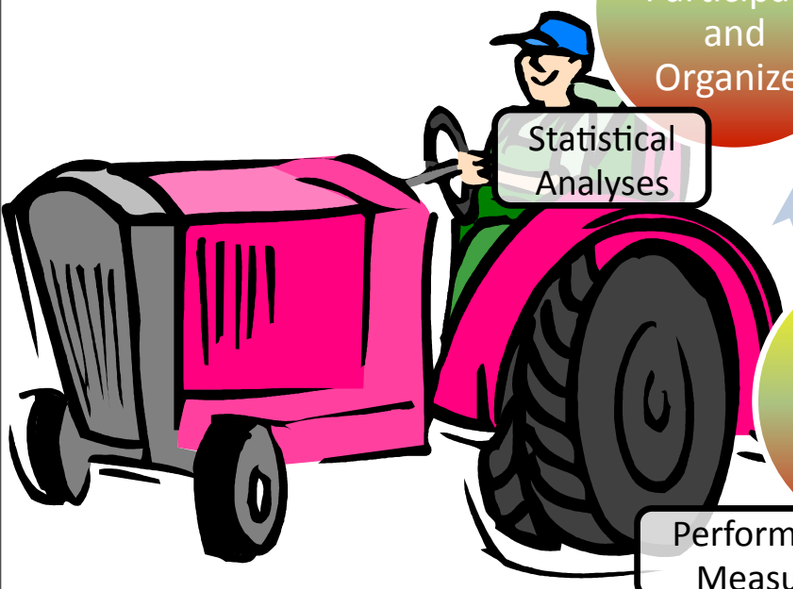
Creation of Pools

Assessor

Relevance Assessment

Visitor, Participant and Organizer

Performance Measures

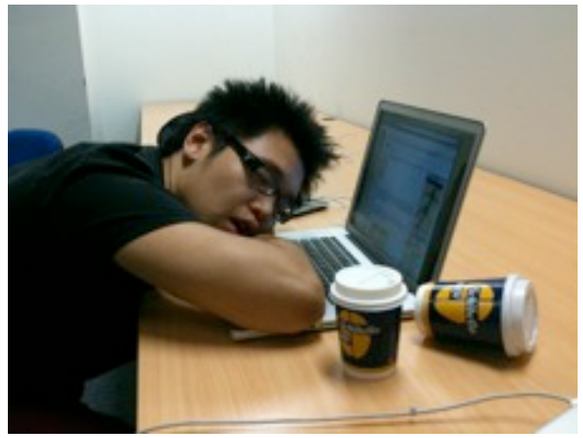


Visitor, Participant and Organizer

Statistical Analyses

How much Valuable are our Data?

- The **TREC 2010 Economic Impact** study estimated in about **30 M$** the **overall investment** in TREC by NIST
  - probably much much more if we had a means to estimate also the investment by participants in TREC
- The **Reliable Information Access (RIA)** workshop in 2004 estimated in **11 to 40 hours per topic** to carry out failure analysis
- They are the **pillars** for all the subsequent **scientific research** and **technology development**
  - TREC estimated the return on investment in the range of 3-5 for each invested dollar



How much Valuable are our Data?

- The **TREC 2010 Economic Impact** study estimated in about **30 M$** the **overall investment** in TREC by NIST

We need:

(i) to guarantee their accessibility and interpretability over time

(ii) to curate and enrich them

(iii) to ease their (re-)use

- TREC estimated the return on investment in the range of 3-5 for each invested dollar

The Format Babele

```
<top>

<num> Number: 303
<title> Hubble Telescope Achievements

<desc> Description:
Identify positive accomplishments of the H
was launched in 1991.

<narr> Narrative:
Documents are relevant that show the Hubbl
new data, better quality data than previou
has increased human knowledge of the unive
to disproving previously existing theories
limited to the shortcomings of the telesco
Details of repairs or modifications to the
reference to positive achievements would r

</top>
```

```
<topic number="6" type="ambiguous">
    <query>
        kcs
    </query>
    <description>
        Find information on the Kansas City Southern railroac
    </description>
    <subtopic number="1" type="nav">
        Find the homepage for the Kansas City Southern railro
    </subtopic>
    <subtopic number="2" type="inf">
        I'm looking for a job with the Kansas City Southern r
    </subtopic>
    <subtopic number="3" type="nav">
        Find the homepage for Kanawha County Schools in West
    </subtopic>
    <subtopic number="4" type="nav">
        Find the homepage for the Knox County School system i
    </subtopic>
    <subtopic number="5" type="inf">
        Find information on KCS Energy, Inc., and their merge
    </subtopic>
</topic>
```

```
<session num="1" starttime="08:59:47.258675">
    <topic>
        <title>
            peacecorp
        </title>
        <desc>
            Find information about the peace corp
        </desc>
        <narr>
            When was it started and by whom? What services does it provide and where does
        </narr>
    </topic>
    <interaction num="1" starttime="09:00:04.155323">
        <query>
            peace corp
        </query>
        <results>
            <result rank="1">
                <url>
                    http://www.peacecorps.gov/
                </url>
                <clueweb09id>
                    clueweb09-en0011-60-08003
                </clueweb09id>
                <title>
                    Peace Corps
                </title>
                <snippet>
                    Fighting hunger, disease, poverty, and lack of opportunity.
                </snippet>
            </result>
            ...
        </results>
        <clicked>
            <click num="1" starttime="09:00:09.943356" endtime="09:01:13.434255">
                <rank>
```

This situation hampers:

- automatic management
- accessibility and interpretability of the data
- ease of (re-)use
- interoperability
- take-up from new comers

The Format Babele

```
303 0 APW19980609.1531 2
303 0 APW19980610.1778 1
303 0 APW19980715.1061 2
303 0 APW19980910.1078 0

1 0 clueweb09-en0120-13-20479 0
1 1 clueweb09-en0120-13-20479 0
1 2 clueweb09-en0120-13-20479 0

101 0 clueweb09-en0047-33-20039 1
101 0 clueweb09-en0004-66-09322 2
101 0 clueweb09-en0033-30-08382 0
101 0 clueweb09-en0000-45-05740 -2
101 0 clueweb09-en0020-92-11795 1

20002 0 clueweb09-en0006-85-33170 1 1 10.5
20004 0 clueweb09-en0005-28-20976 1 1 10.5
20006 0 clueweb09-en0010-07-21538 1 1 10.5
```

ad-hoc

diversity

ad-hoc with grades

relevance feedback

This situation hampers:

- automatic management
- accessibility and interpretability of the data
- ease of (re-)use
- interoperability
- take-up from new comers

The Format Babele

We need:

(i) to agree on a common data model which allows for extension

(ii) to provide the basic experimental data with proper metadata (descriptive, administrative, copyright, ...)

- interoperability
- take-up from new comers

Referenceability and Traceability

- Explanation of experimental data is usually reported in scientific papers that do not provide direct links to them
- It is often difficult to exactly know which data have been used in a paper and have access to them
- It is ever more difficult to exactly know and reproduce data cleaning and processing operations

Referenceability and Traceability

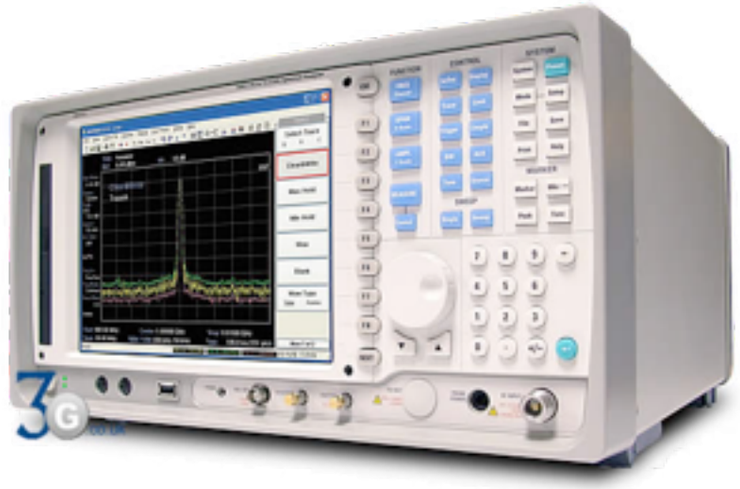
We need:

(i) to have the possibility of citing experimental data in our papers as any other references

(ii) to make our papers actionable and executable providing access to the mentioned experimental data

Automation Increase

- Automatic evaluation of a system
  - increase in the number of experiment
  - more detailed information
  - reduction of the effort to evaluate a system
  - more comparability
- Programmatic access
  - increase in the number of experiment
  - new usages and new applications driven by the community
  - evaluation as part of the development process, like unit testing
  - opportunistic use by third parties with confidentiality requirements
- Improve ground-truth creation
  - reduction of human effort
  - opportunity for more user-centred evaluation



Goals

- Lowering the effort (and costs) for carrying out experimental evaluation
- Reducing fragmentation in the experimental evaluation (diverse tasks and metrics, heterogeneous collections, different systems, ...)
- Providing management, curation, accessibility over the time to experimental data as well richer interaction, visualization, and exploration of them
- Supporting continuous evaluation of information access and retrieval systems
- Offering programmatic access to the experimental data and extensibility

Bridging between IR and DB

How to Proceed?

- Modelling

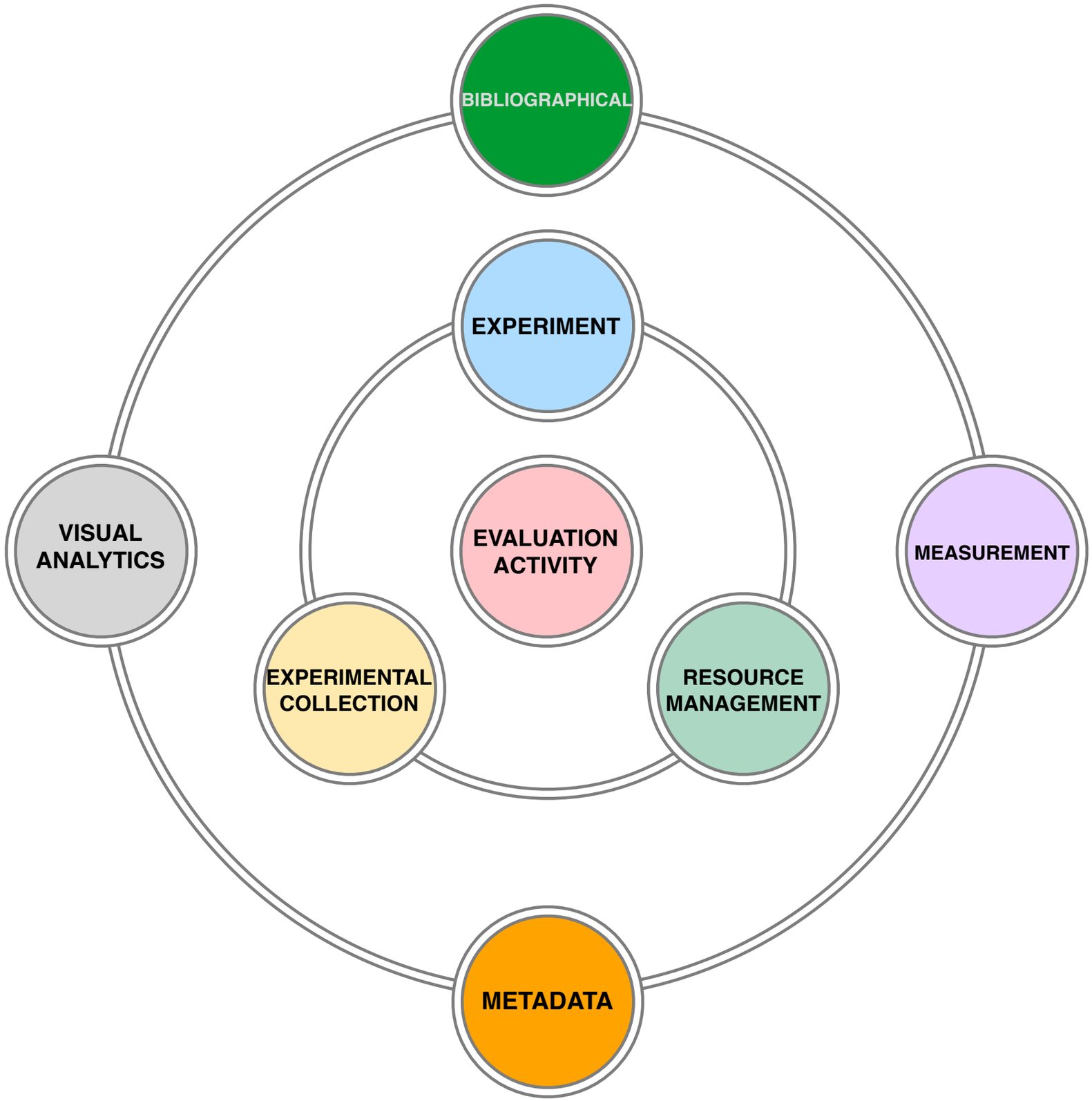
- Designing

- Developing
- Sharing

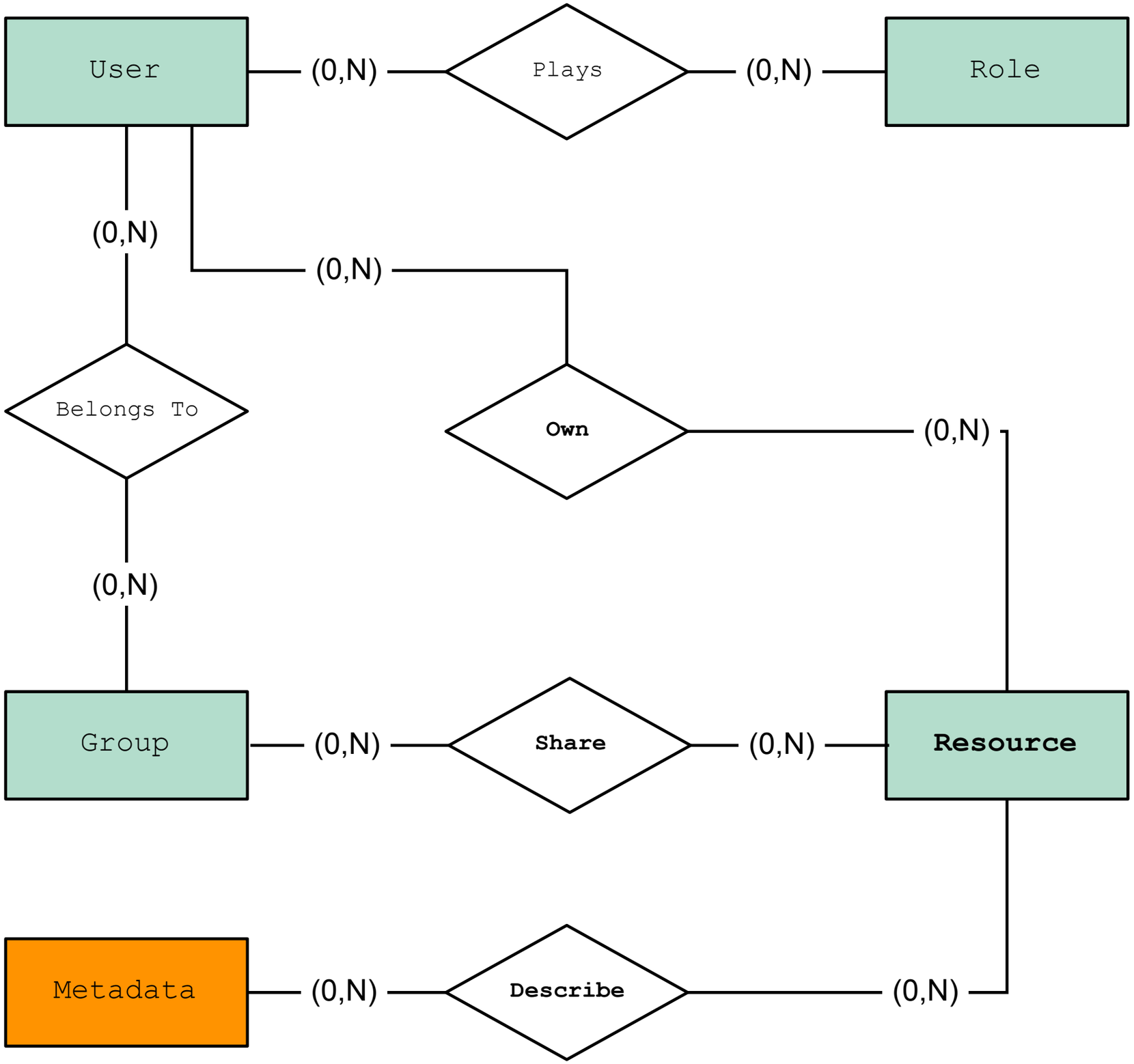
- Extending

Modelling

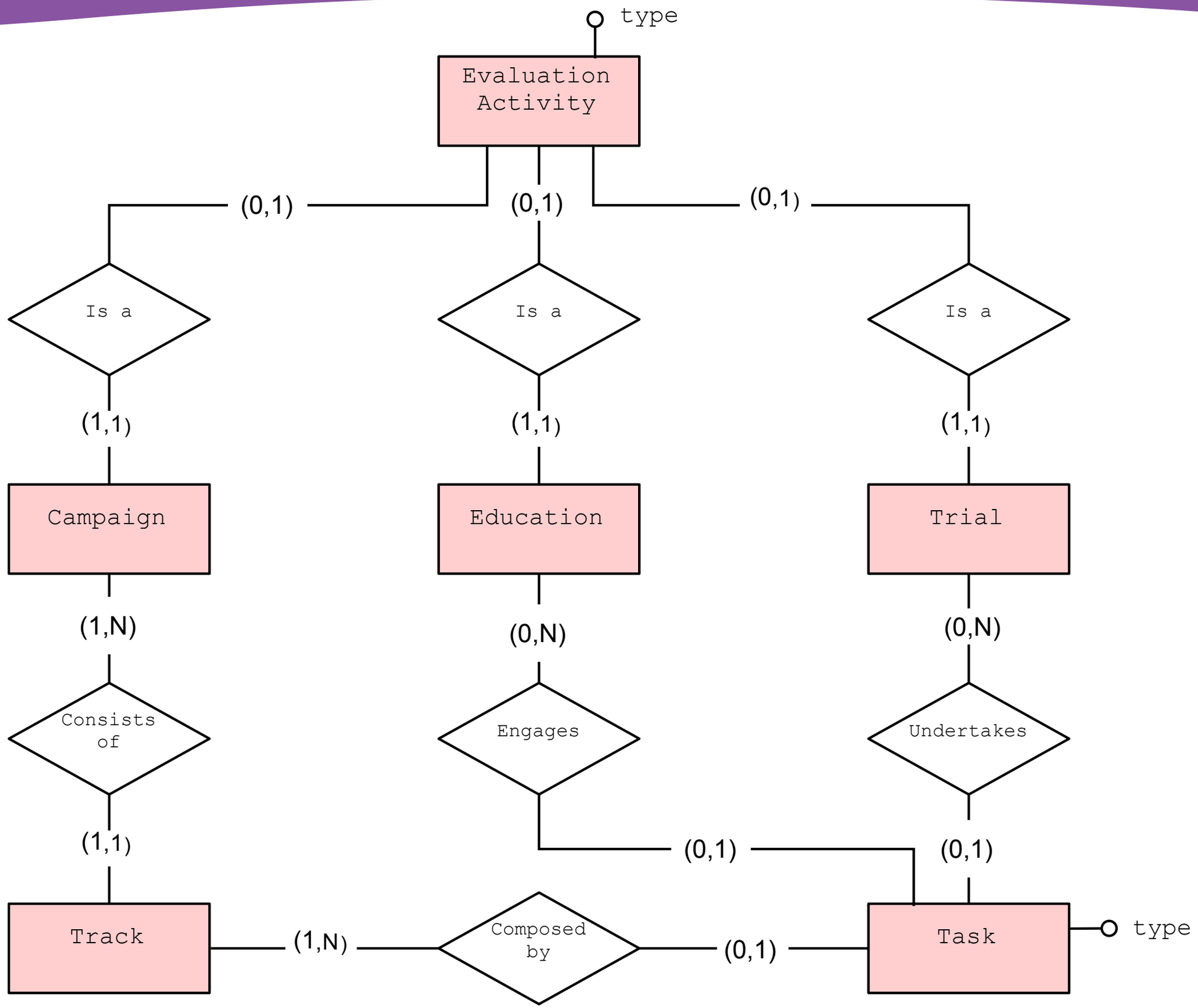
Modelling Areas



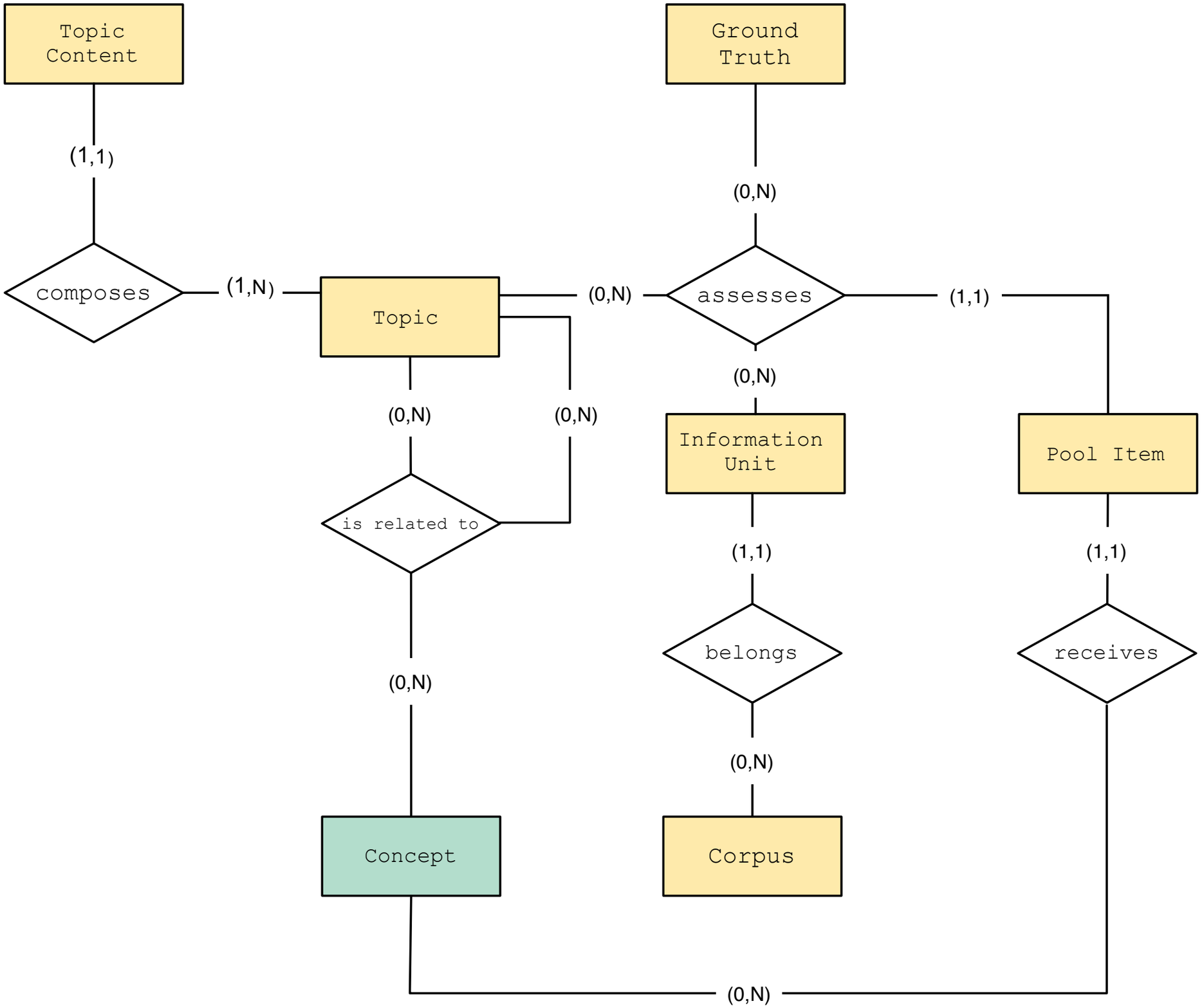
Resource Management and Metadata



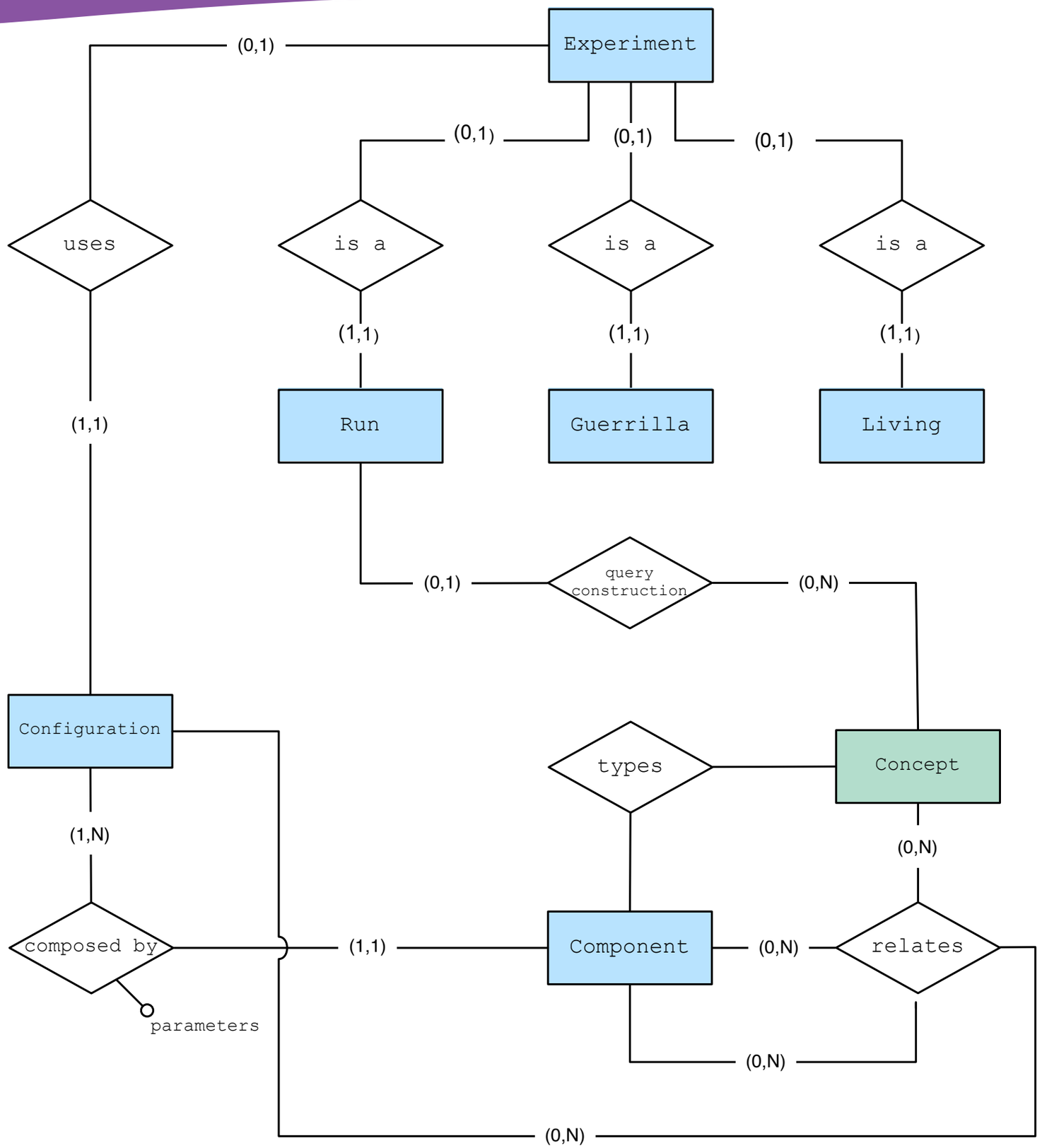
Evaluation activities



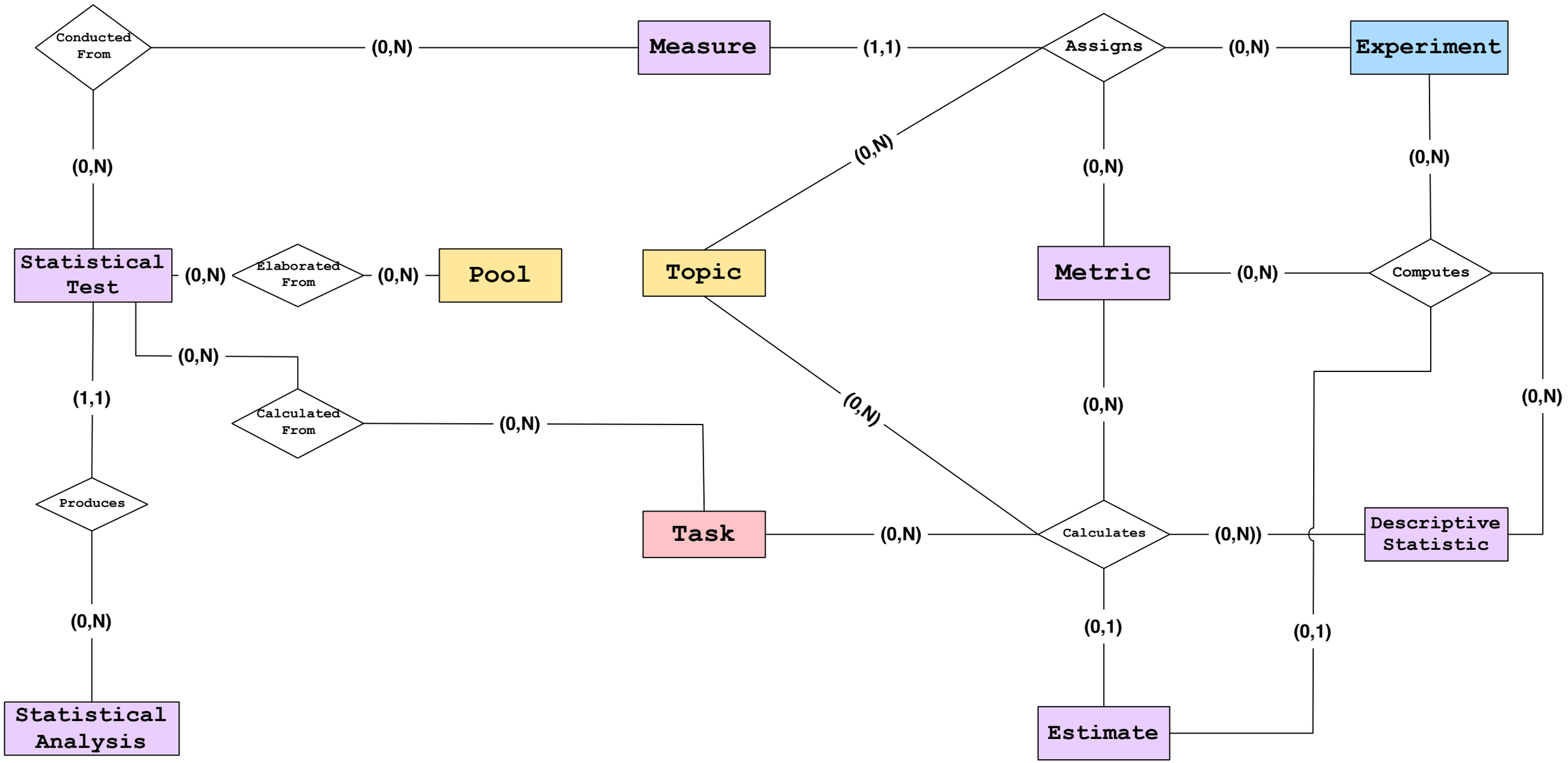
Experimental Collections



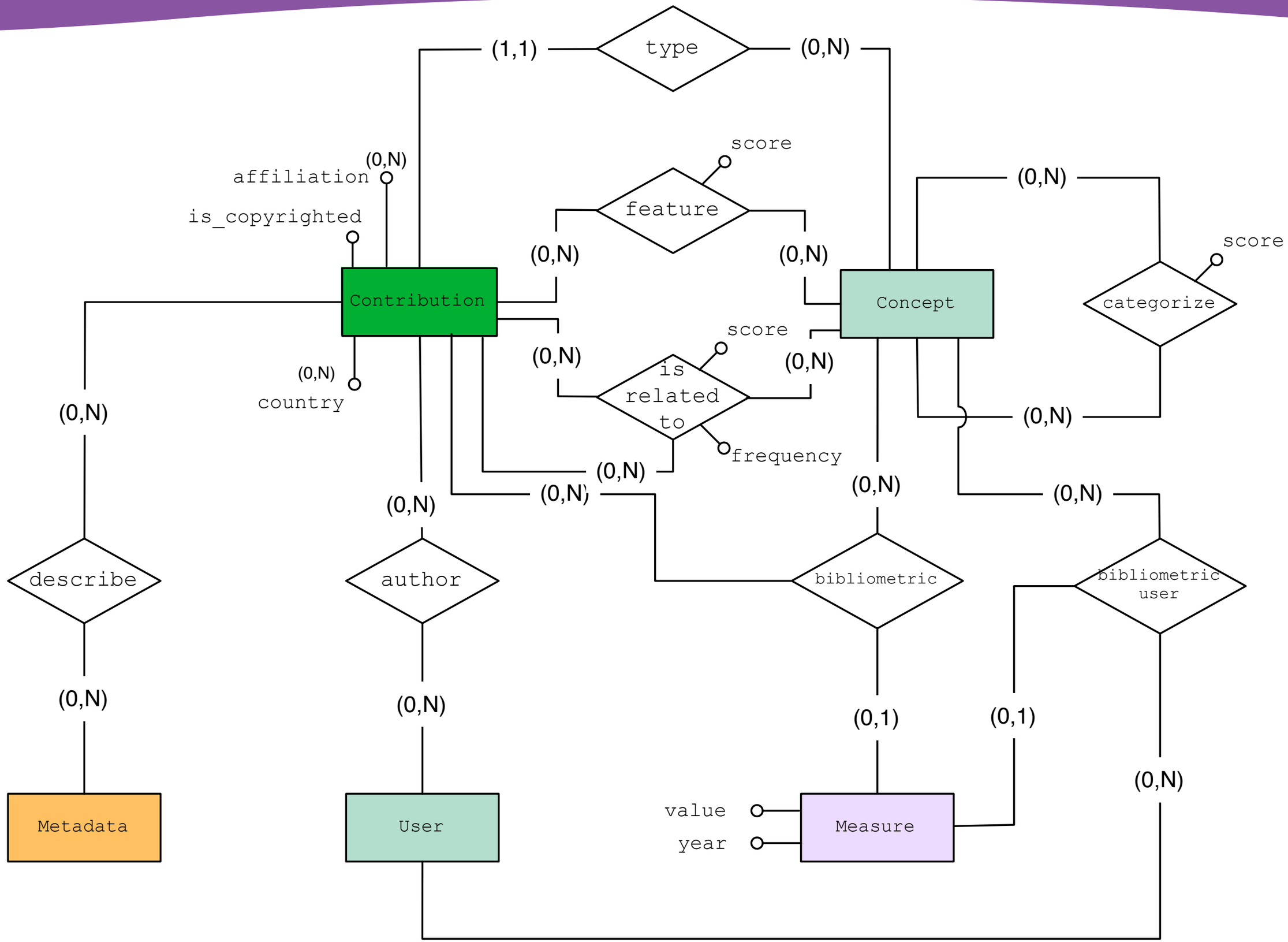
Experiments



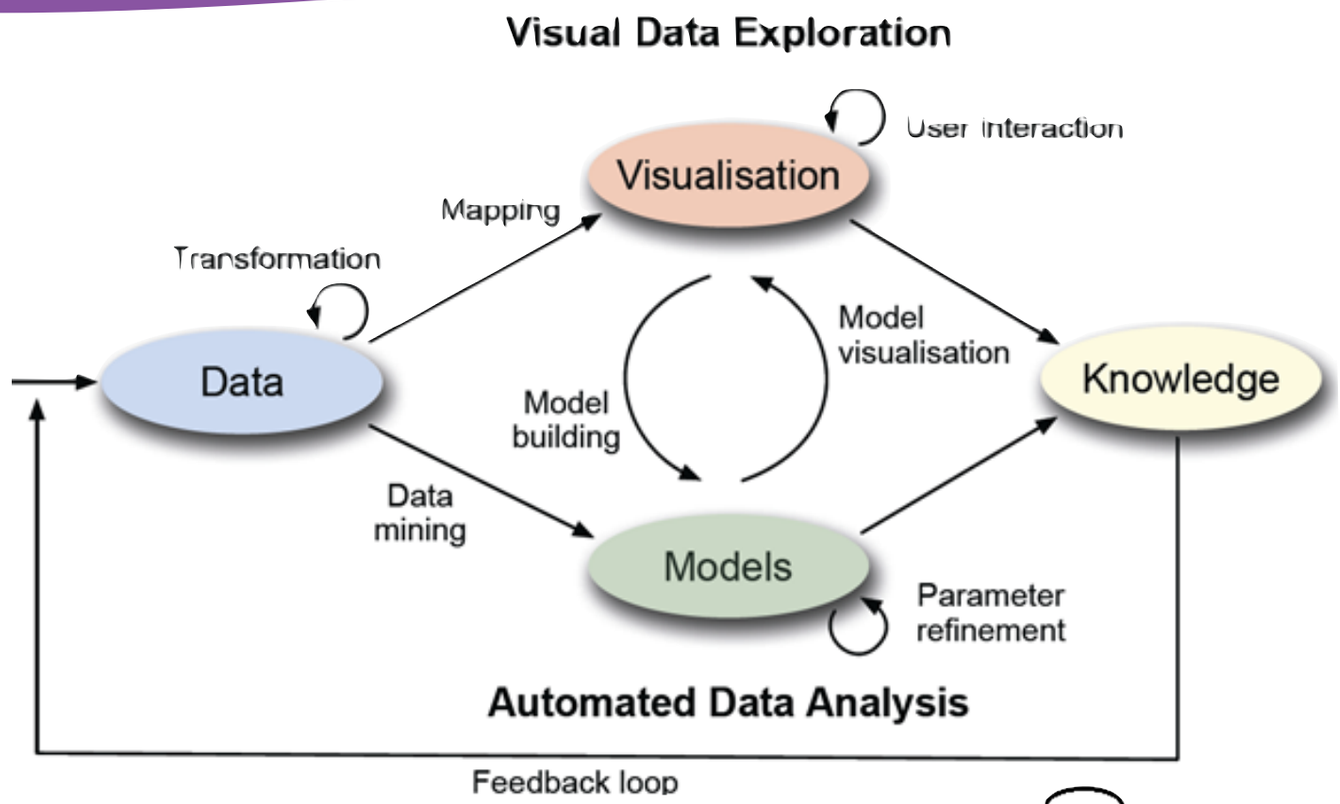
Measurement

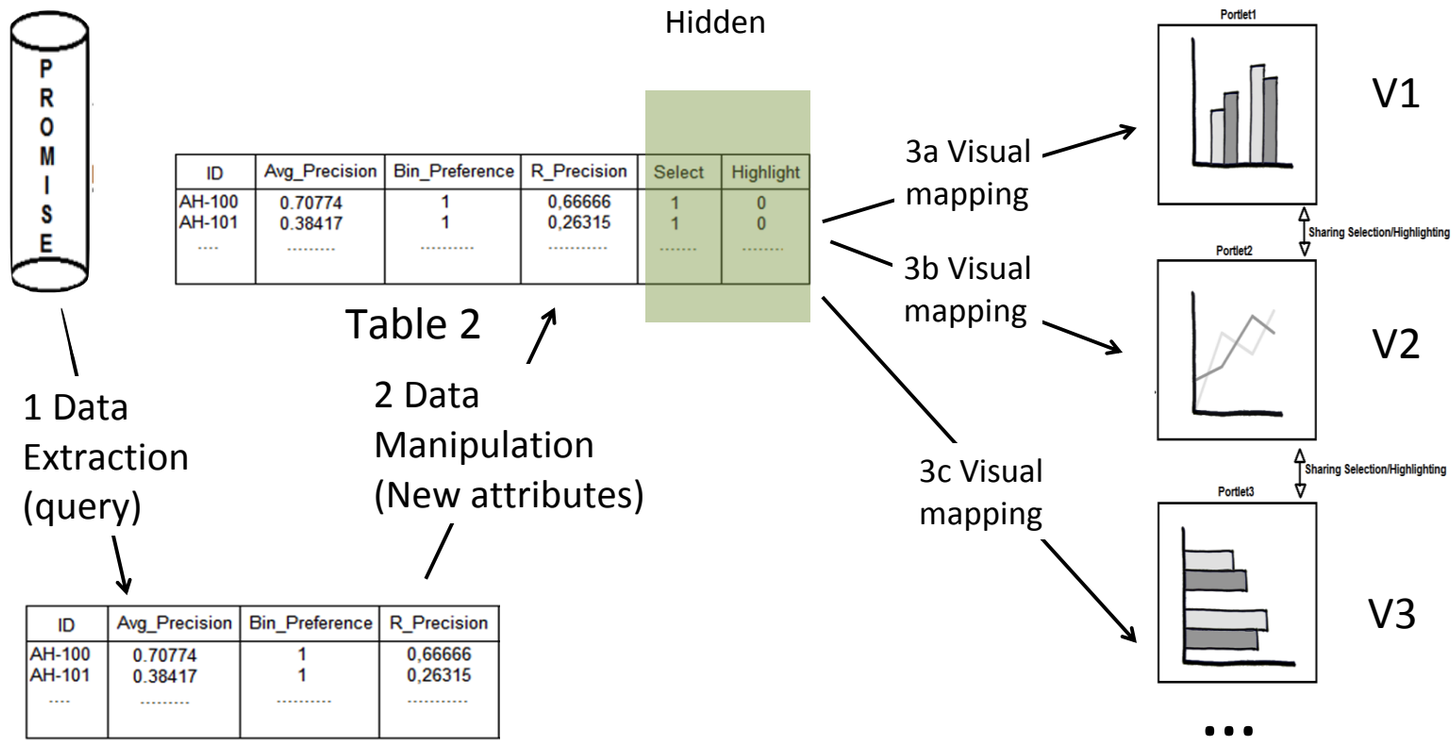


Bibliography and Impact Analysis



Visual Analytics for IR Experimental Evaluation





Visual Analytics for IR Experimental Evaluation

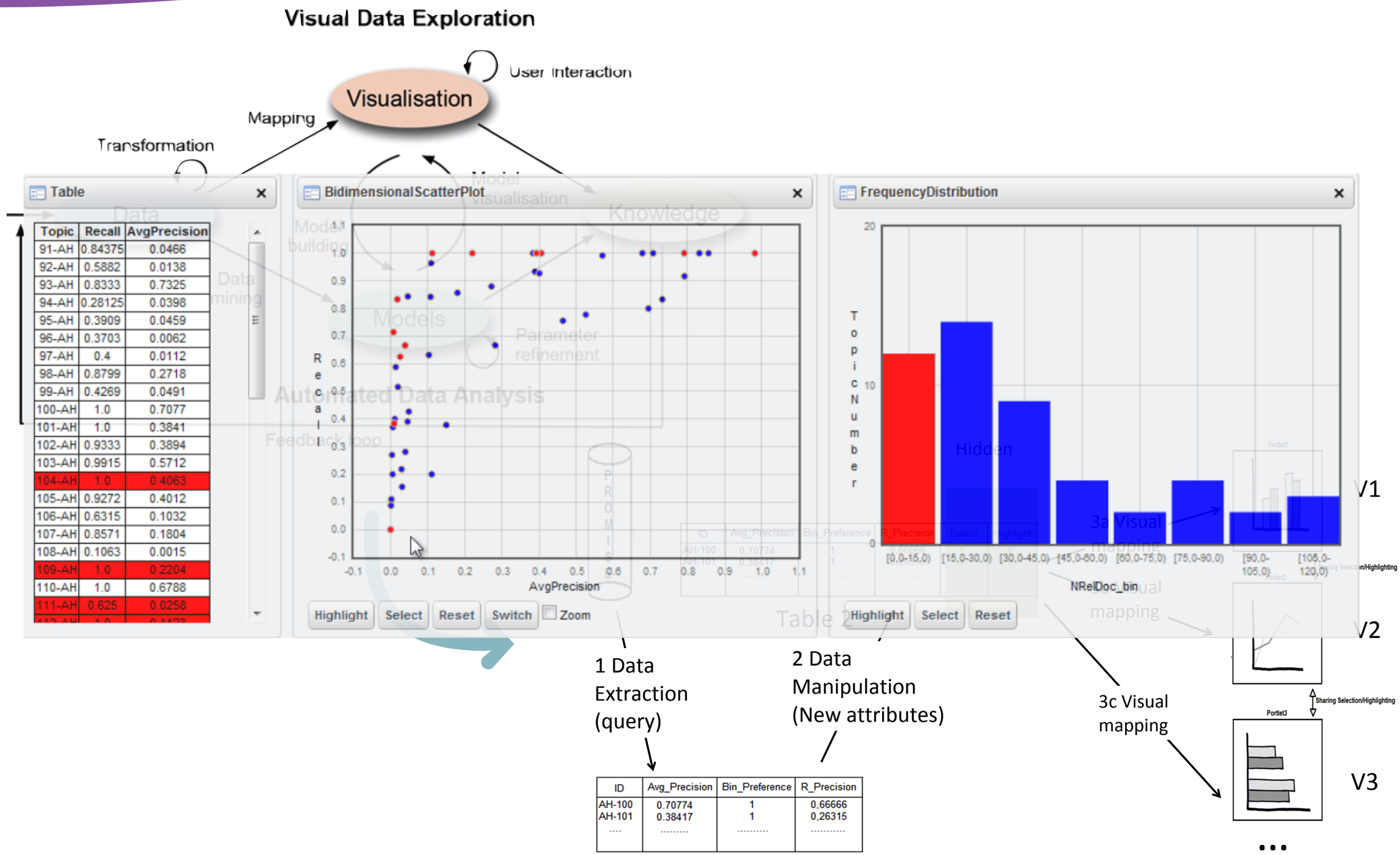
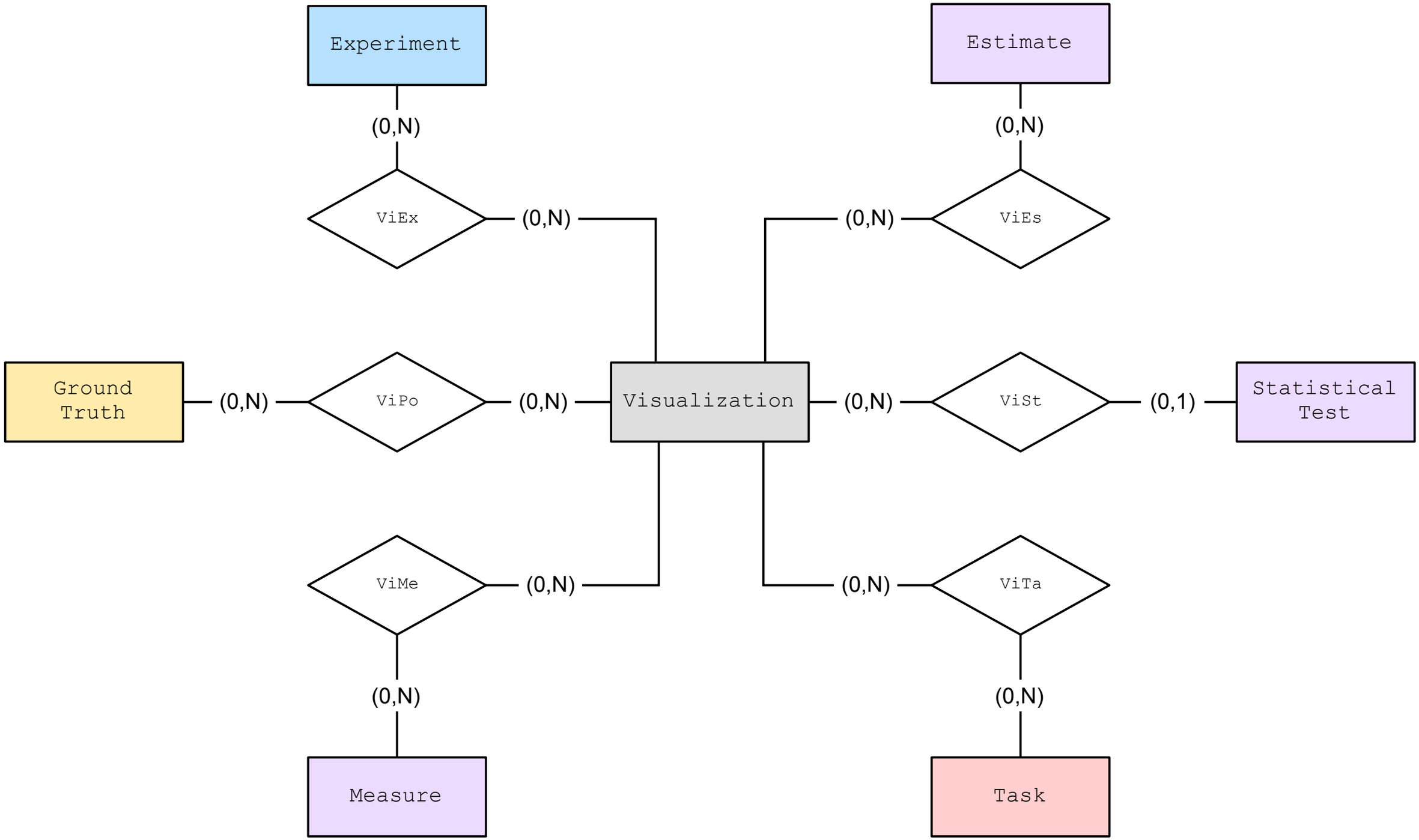


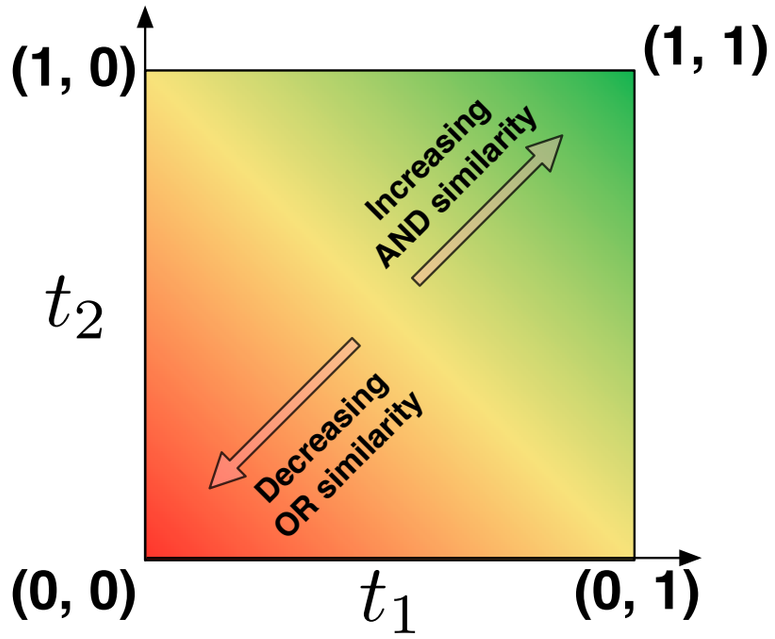
Table 1

Visual Analytics



Search Model

- Extended boolean retrieval model for supporting exact and best match queries on the managed resources
- P-norm is used to compute distance

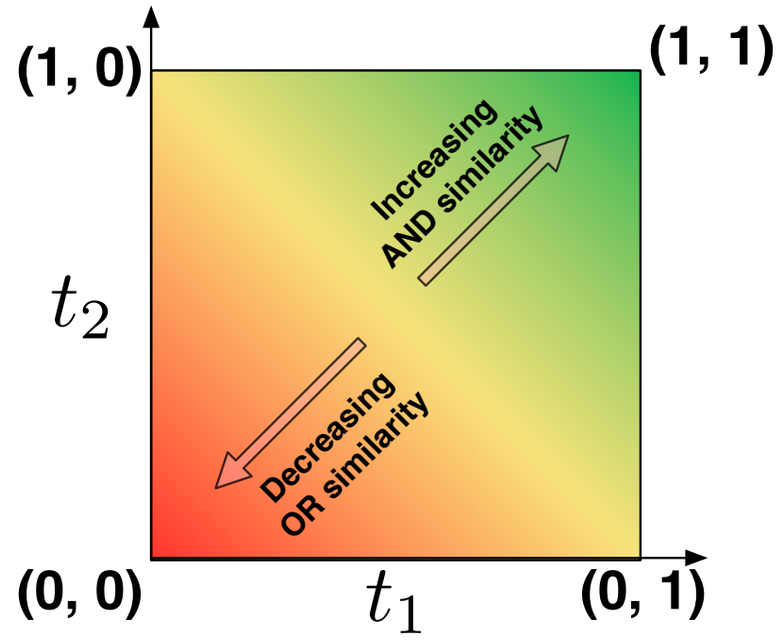


$$\mathrm{sim}_p^{\mathbf{or}}(r, q) = \left[\frac{\mathrm{sim}(r, t_1)^p + \cdots + \mathrm{sim}(r, t_n)^p}{n}\right]^{\frac{1}{p}}$$

$$\mathrm{sim}_p^{\mathbf{and}}(r, q) = 1 - \left[\frac{(1 - \mathrm{sim}(r, t_1))^p + \cdots + (1 - \mathrm{sim}(r, t_n))^p}{n}\right]^{\frac{1}{p}}$$

Search Model

- Extended boolean retrieval model for supporting exact and best match queries on the managed resources
- P-norm is used to compute distance

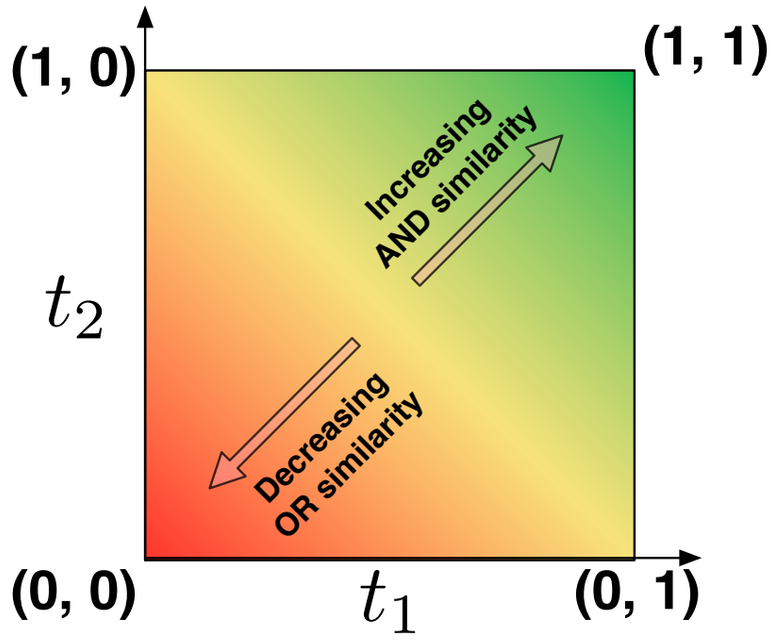


- Best match (p = 1)

$$\mathrm{sim}^{\mathbf{or}}_{\mathtt{best}}(r,q) = \mathrm{sim}^{\mathbf{and}}_{\mathtt{best}}(r,q) = \frac{\mathrm{sim}(r,t_1) + \mathrm{sim}(r,t_2) + \cdots + \mathrm{sim}(r,t_n)}{n}$$

Search Model

- Extended boolean retrieval model for supporting exact and best match queries on the managed resources
- P-norm is used to compute distance



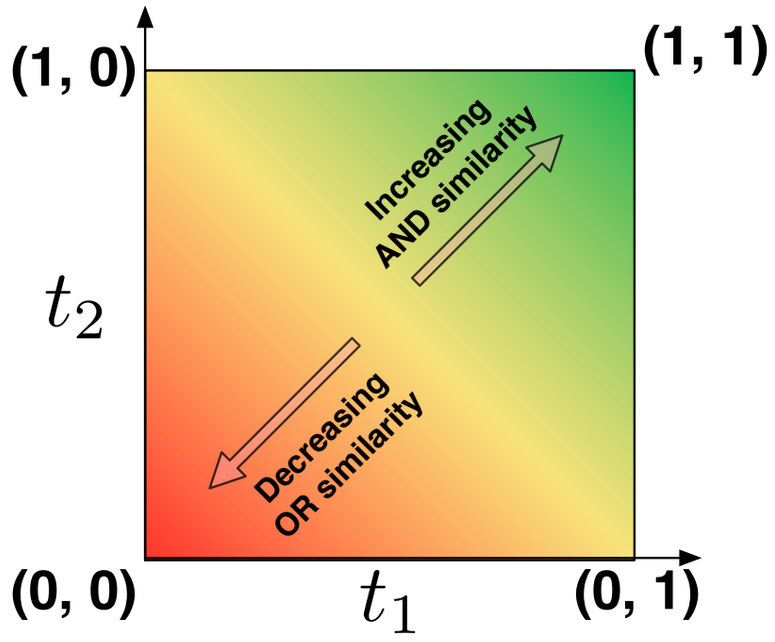
- Loose match (p = 5)

$$\mathrm{sim}^{\mathbf{or}}_{\texttt{loose}}(r,q) = \sqrt[5]{\frac{\mathrm{sim}^5(r,t_1) + \mathrm{sim}^5(r,t_2) + \cdots + \mathrm{sim}^5(r,t_n)}{n}}$$

$$\mathrm{sim}^{\mathbf{and}}_{\texttt{loose}}(r,q) = 1 - \sqrt[5]{\frac{\big(1-\mathrm{sim}(r,t_1)\big)^5 + \big(1-\mathrm{sim}(r,t_2)\big)^5 + \cdots + \big(1-\mathrm{sim}(r,t_n)\big)^5}{n}}$$

Search Model

- Extended boolean retrieval model for supporting exact and best match queries on the managed resources
- P-norm is used to compute distance



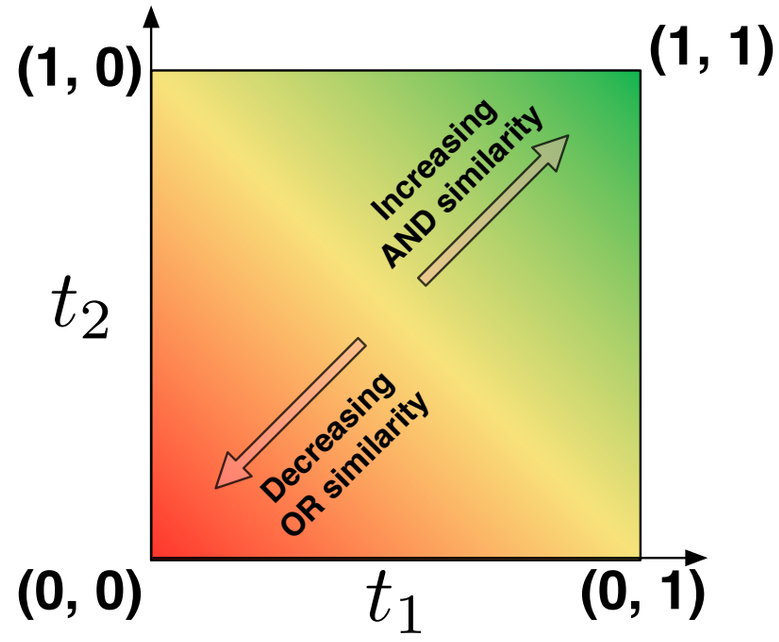
- Fuzzy match (p = 2)

$$\mathrm{sim}^{\mathbf{or}}_{\texttt{fuzzy}}(r,q) = \sqrt[2]{\frac{\mathrm{sim}^2(r,t_1) + \mathrm{sim}^2(r,t_2) + \cdots + \mathrm{sim}^2(r,t_n)}{n}}$$

$$\mathrm{sim}^{\mathbf{and}}_{\texttt{fuzzy}}(r,q) = 1 - \sqrt[2]{\frac{\big(1-\mathrm{sim}(r,t_1)\big)^2 + \big(1-\mathrm{sim}(r,t_2)\big)^2 + \cdots + \big(1-\mathrm{sim}(r,t_n)\big)^2}{n}}$$

Search Model

- Extended boolean retrieval model for supporting exact and best match queries on the managed resources
- P-norm is used to compute distance



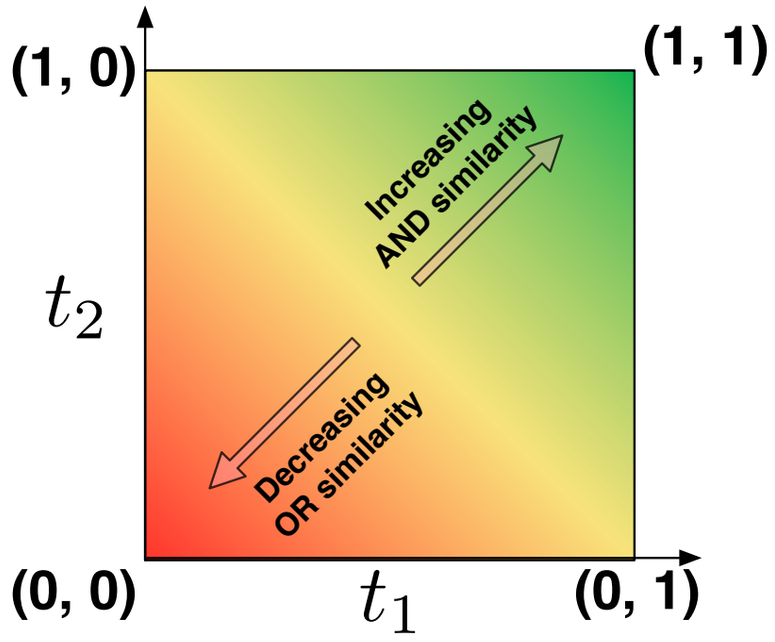
- Exact match (p = ∞)

$$\mathrm{sim}^{\mathbf{or}}_{\mathtt{exact}}(r, q) = \max\left(\mathrm{sim}(r, t_1), \mathrm{sim}(r, t_2), \ldots, \mathrm{sim}(r, t_n)\right)$$

$$\mathrm{sim}^{\mathbf{and}}_{\mathtt{exact}}(r, q) = \min\left(\mathrm{sim}(r, t_1), \mathrm{sim}(r, t_2), \ldots, \mathrm{sim}(r, t_n)\right)$$

Search Model

- Extended boolean retrieval model for supporting exact and best match queries on the managed resources
- P-norm is used to compute distance

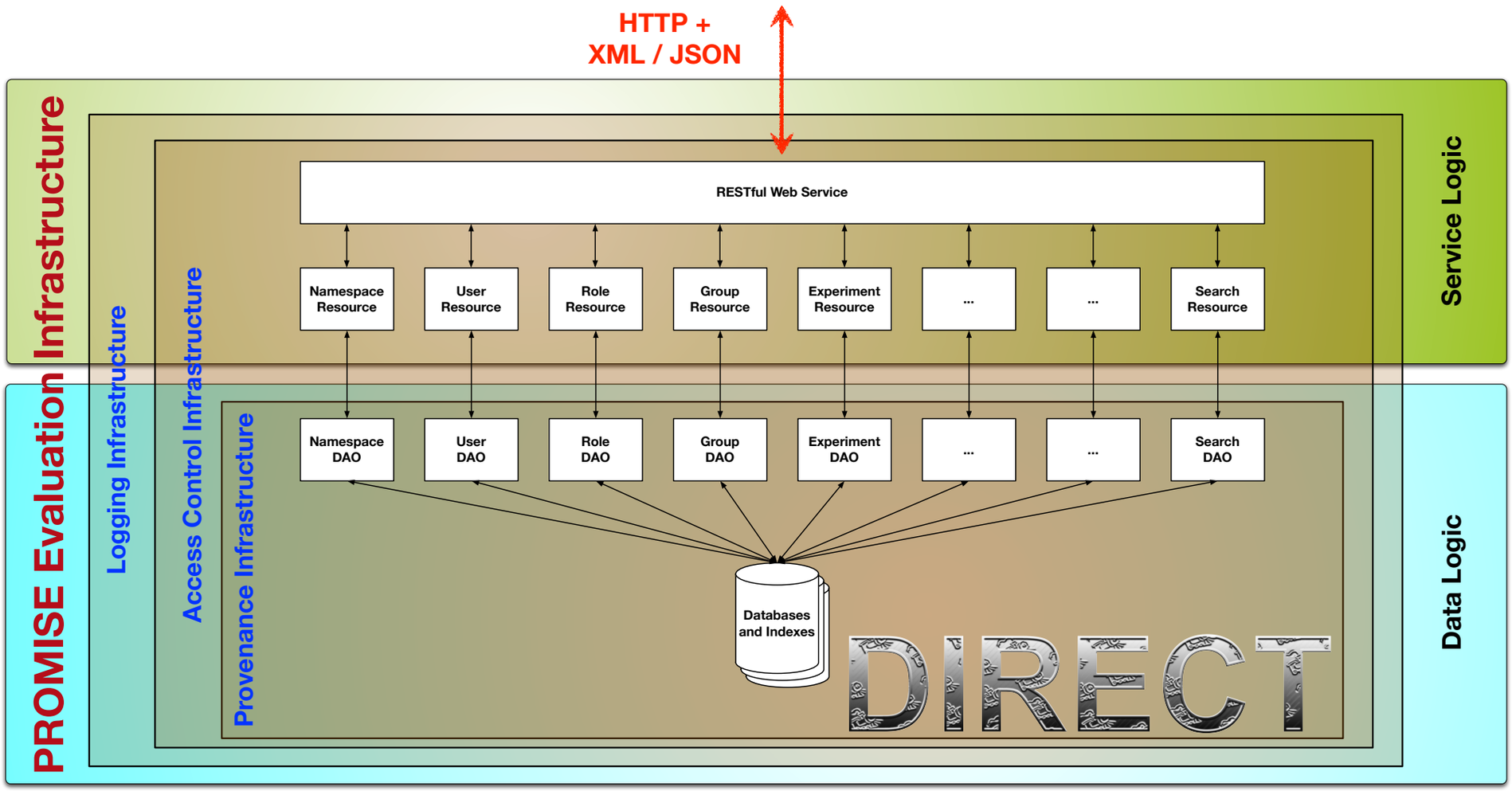


- Query language based on the CQL (Contextual Query Language) syntax
  - It allows for embedding queries in URL and linking to them

```
direct.experiment.description = "language model"
                and/match=exact
        direct.experiment.user.id = unine
```

Designing and Developing

Architecture



Architecture

Evaluation Infrastructure Portal

Application + Interface Logic

Experiment Submission Portlet

Topic Creation Portlet

DIRECT Web Application

Parallel Coordinated View Portlet

Rank Analysis Portlet

Visual Analytics Environment

Portlet

Filtered News Portlet

Collaborative Enviroment

Portlet

Portlet

...

PROMISE Evaluation Infrastructure

Logging Infrastructure

Access Control Infrastructure

Provenance Infrastructure

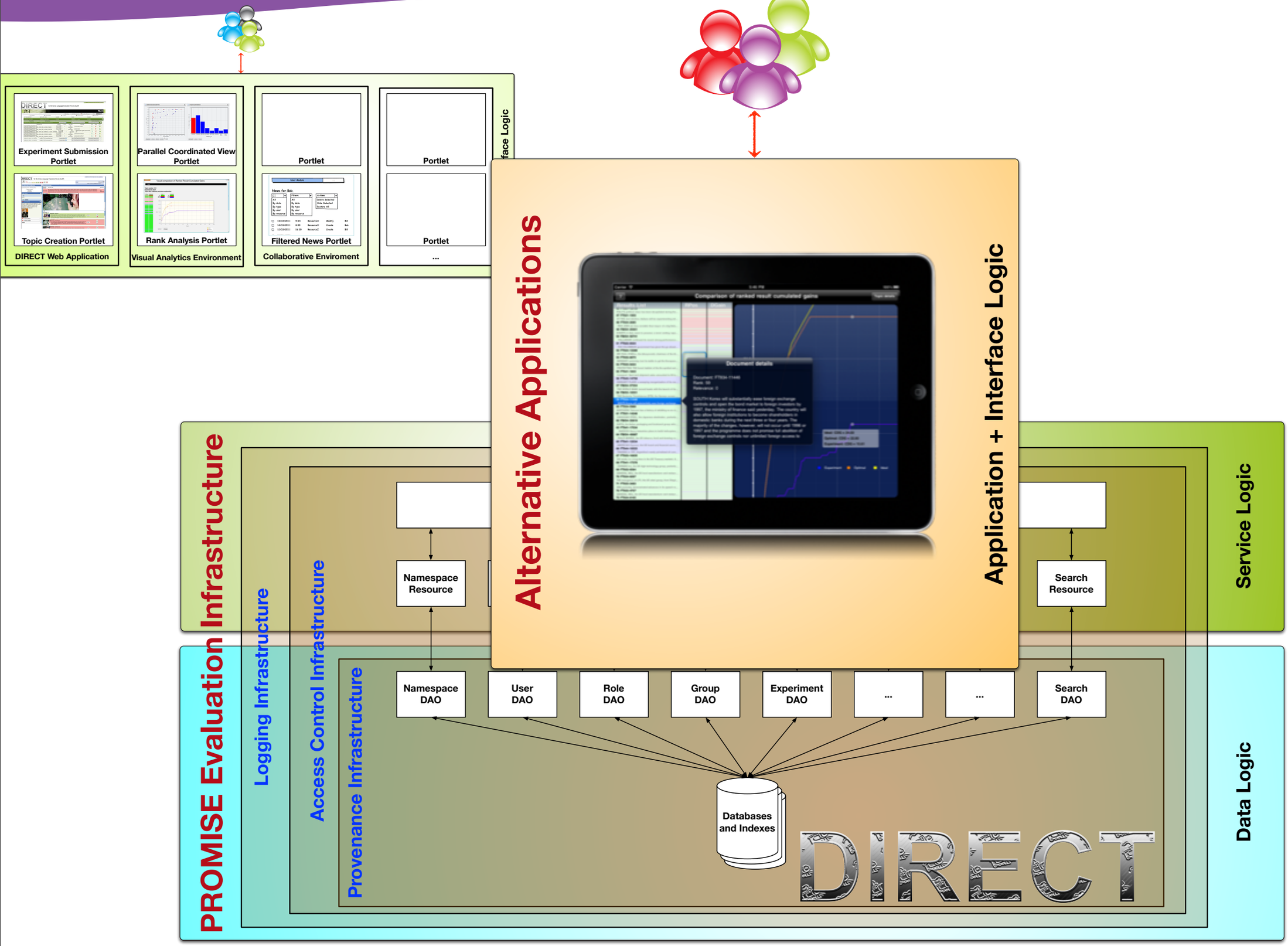
Namespace DAO | User DAO | Role DAO | Group DAO | Experiment DAO | ... | ... | Search DAO

Databases and Indexes

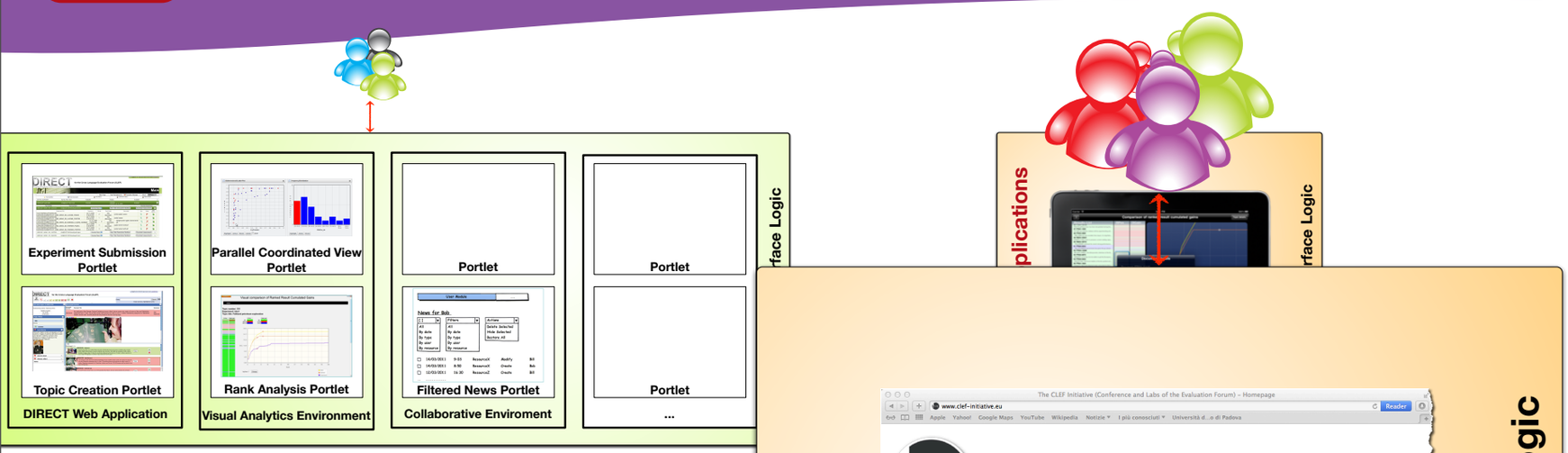
DIRECT

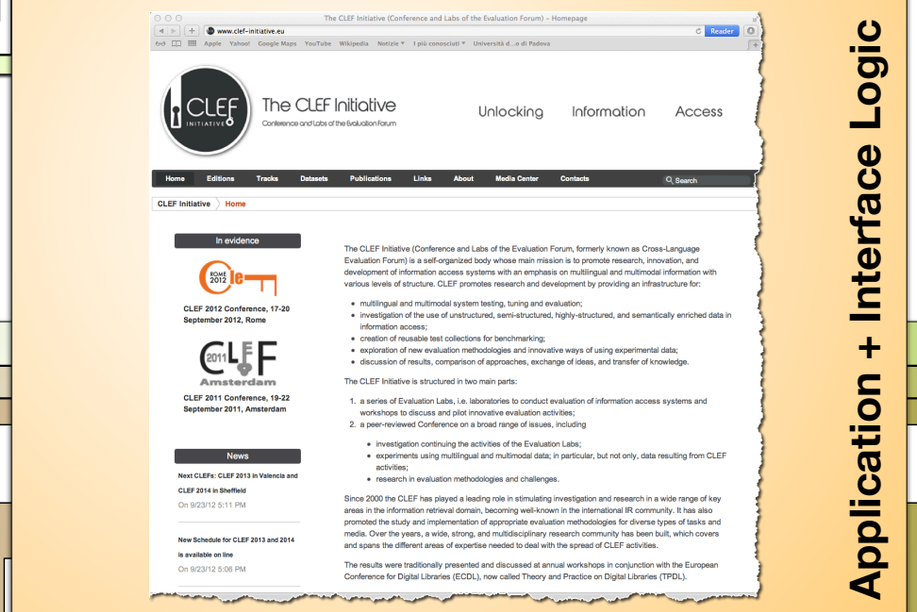
Data Logic

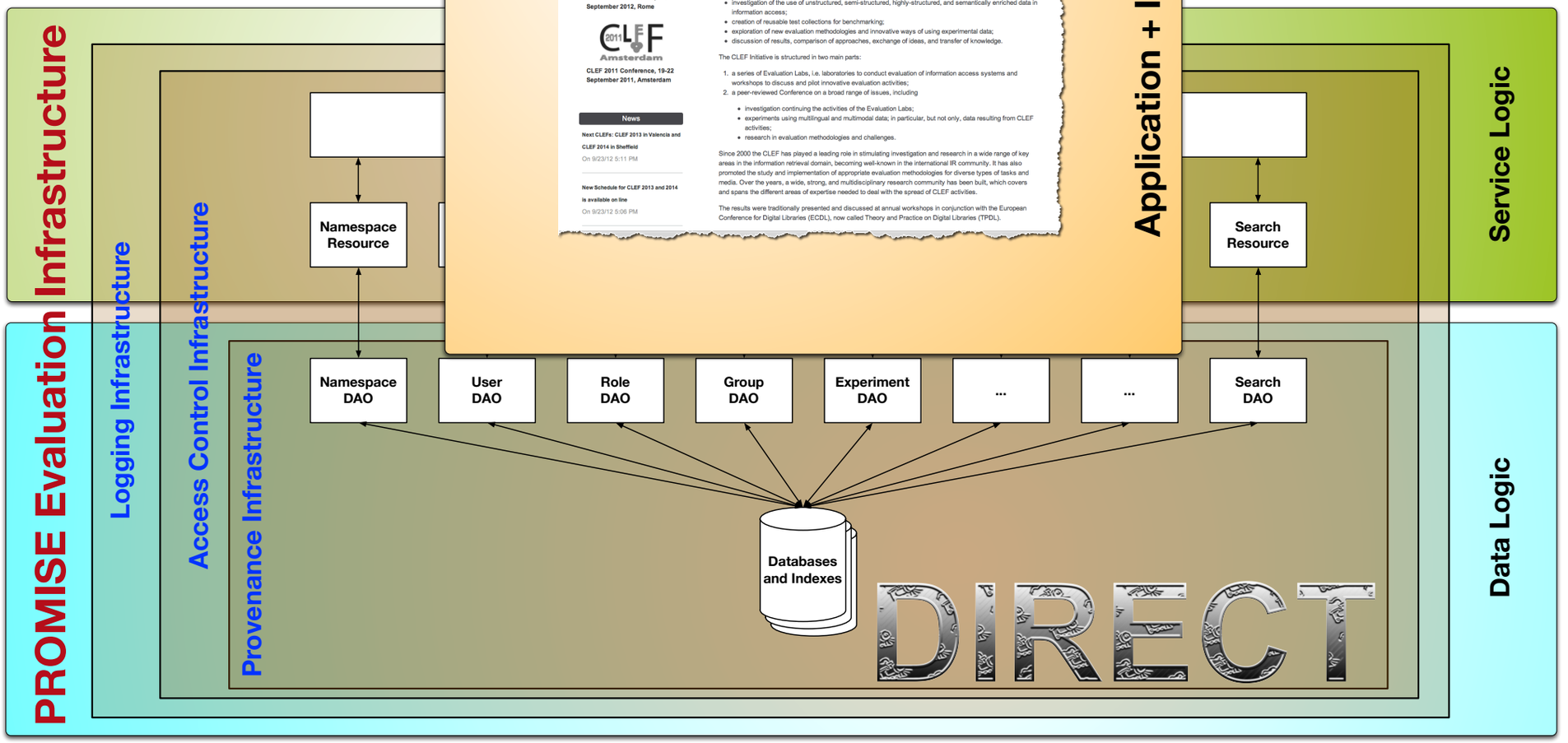
Architecture



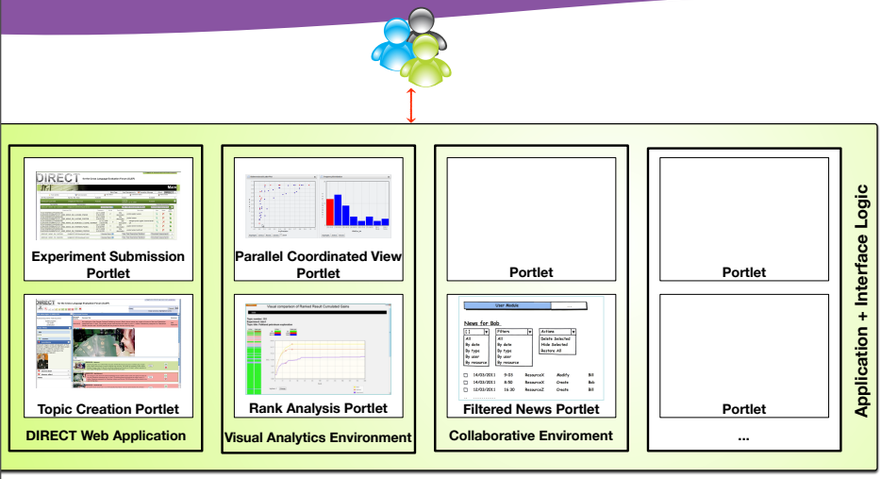
Architecture

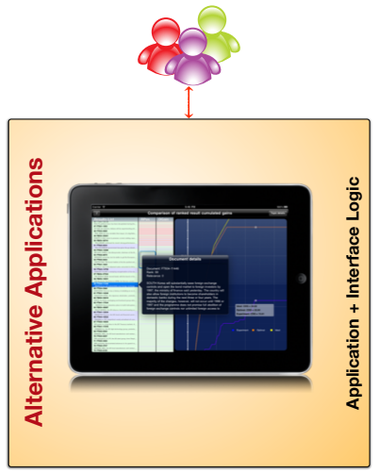


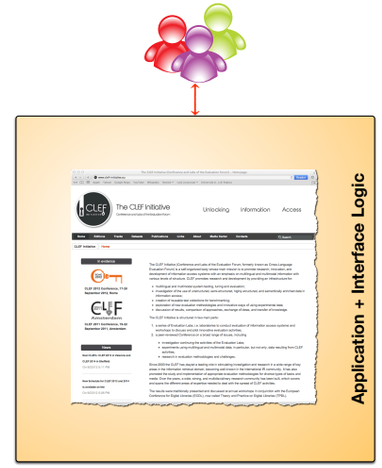


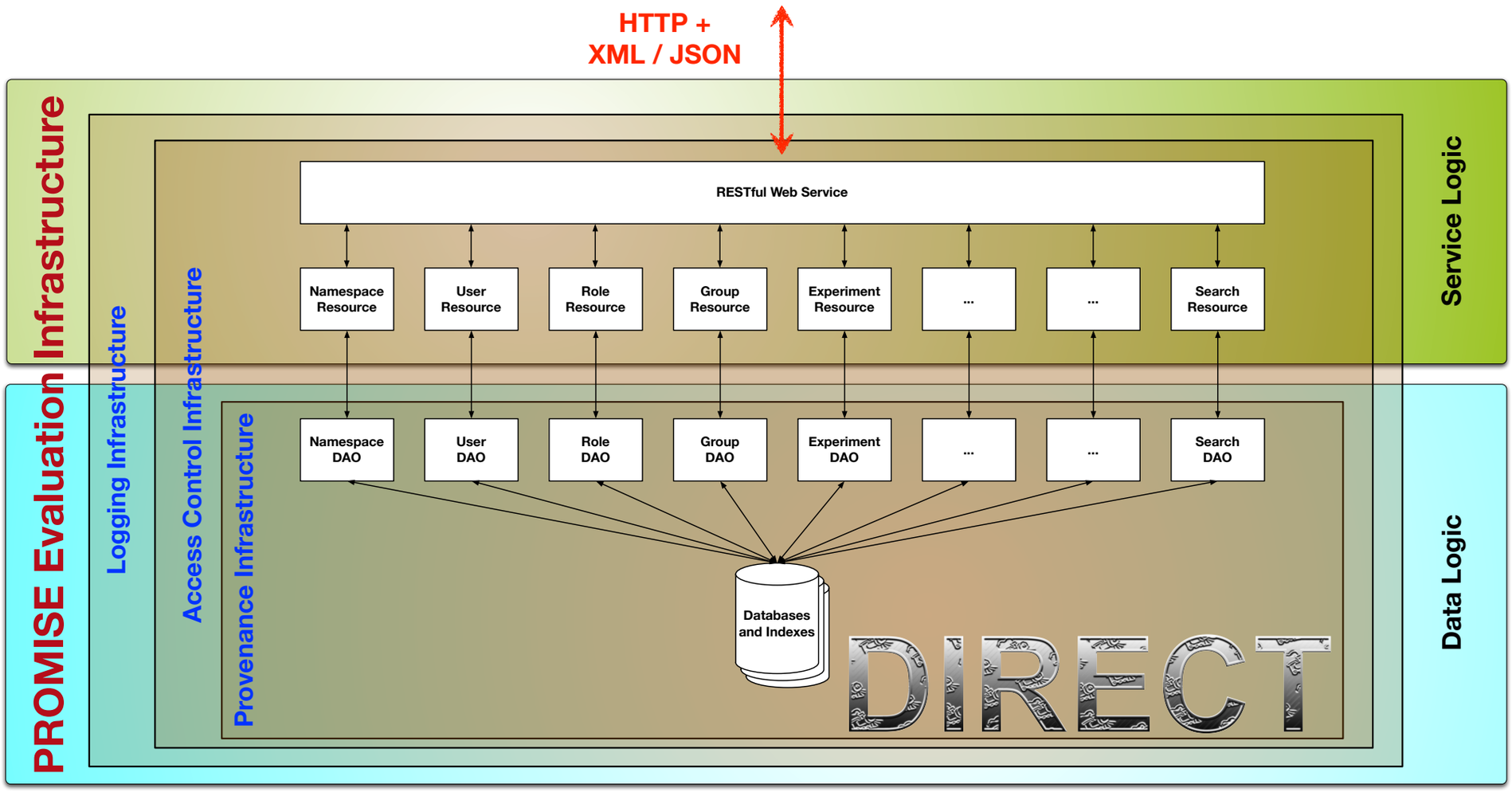


Architecture









The DIRECT Application

Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) – Portal Main Page

direct.dei.unipd.it/public/portal.tiles

Logged as clef@tu-chemnitz.de @ CLEF2012 [Logout] - Monday 17 December 2012

DIRECT

Distributed Information Retrieval Evaluation Campaign Tool

for Evaluation of Multilingual Information Access in PROMISE NoE

Portal Main Page

Translation Manager

Main Page

DIRECT Portal

About DIRECT

User Management

- Campaigns
  - CLEF 2000
  - CLEF 2001
  - CLEF 2002
  - CLEF 2003
  - CLEF 2004
  - CLEF 2005
  - CLEF 2006
  - CLEF 2007
  - CLEF 2008
  - CLEF 2009
    - Tracks
      - Ad-Hoc Persian Track
        - Tasks
          - Ad-Hoc TEL Bilingual Persian Task
          - Ad-Hoc TEL Monolingual Persian Task
            - Download Pool
            - Download Topics
      - Ad-Hoc Persian Topic Creation Track
      - Ad-Hoc Robust Track
      - Ad-Hoc TEL Track
      - Ad-Hoc TEL Topic Creation Track
      - Grid@CLEF Pilot Track
      - ImageCLEF Track
  - CLEF Test

Select all | Unselect all | Download Selected

< prev (1 of 1) next > Show 20 rows

	Identifier	Participant	Description	Query Construction	Source Language	Is Pooled	View	Download
☐	AH-PERSIAN-MONO-FA-CLEF2009.JHU-APL.JHUFA4R100TD	jhu-apl	4-grams; 100 rf terms	AUTOMATIC	fa	true		
☐	AH-PERSIAN-MONO-FA-CLEF2009.JHU-APL.JHUFA4R100TDN	jhu-apl	4-grams; 100 rf terms	AUTOMATIC	fa	true		
☐	AH-PERSIAN-MONO-FA-CLEF2009.JHU-APL.JHUFA5R100TD	jhu-apl	5-grams; RF 100 terms	AUTOMATIC	fa	true		
☐	AH-PERSIAN-MONO-FA-CLEF2009.JHU-APL.JHUFASK41R400TD	jhu-apl	skip 4-grams (1 skip); 400 RF terms	AUTOMATIC	fa	true		
☐	AH-PERSIAN-MONO-FA-CLEF2009.JHU-APL.JHUFATR5R50TD	jhu-apl	truncated word forms; 50 RF terms	AUTOMATIC	fa	true		
☐	AH-PERSIAN-MONO-FA-CLEF2009.OPENTEXT.OTFA09T	opentext	title-only, Neuchatel stemmer	AUTOMATIC	fa	true		
☐	AH-PERSIAN-MONO-FA-CLEF2009.OPENTEXT.OTFA09TD	opentext	same as otFA09t except both title and description fields used	AUTOMATIC	fa	true		
☐	AH-PERSIAN-MONO-FA-CLEF2009.OPENTEXT.OTFA09TDE	opentext	blind feedback based on top-3 rows of otFA09td	AUTOMATIC	fa	true		
☐	AH-PERSIAN-MONO-FA-CLEF2009.OPENTEXT.OTFA09TDZ	opentext	depth-10000 sampling run (orig ranks in first 5 decimal places of rsv)	AUTOMATIC	fa	true		
☐	AH-PERSIAN-MONO-FA-CLEF2009.QAZVINIAU.IAUPERFA1	qazviniau	(Title+Desc), Stemmed Collection, Stemmed Queries, Query Stops, PRF(5,10)	AUTOMATIC	fa	true		

Batch Processing via Scripts

```perl
while (my $ranking = shift) {
    $rankname[$i] = $ranking;
    my $runcount;

    open(IN, $ranking) or die "Can't read $ranking: $1\n";
    while (<IN>) {

#       [...]

    $rank[$i]{$run} = $score;
    }
    close(IN);
}

# [...]

for $i (0 .. $numrank - 2) {
    for my $j ($i + 1 .. $numrank - 1) {
    my $t = tau($i, $j);
    print "$t\n";
    $sum += $t;

#   [...]

}

sub tau {
    my ($i, $j) = @_;
    my ($con, $dis, $tie_i, $tie_j, $n, @pairs);
    my (@runs, %runstat);

    for my $run (sort { $rank[$i]{$b} <=> $rank[$i]{$a} }
         keys %{ $rank[$i] }) {
    push @pairs, [ $rank[$i]{$run}, $rank[$j]{$run} ];
    push @runs, $run;
    }

#   [...]

    return $t;
}
```

```perl
while (my $ranking = shift) {
    $rankname[$i] = $ranking;
    my $runcount;

    my $ua = LWP::UserAgent->new();
    my $response = $ua->get(sprintf(
        'http://direct.dei.unipd.it/run/' . $ranking . '/items',
        uri_escape($text),
        uri_escape($key),
    ));

    while (<response>) {

#       [...]

    $rank[$i]{$run} = $score;
    }

}

# [...]

for $i (0 .. $numrank - 2) {
    for my $j ($i + 1 .. $numrank - 1) {
    my $t = tau($i, $j);
    print "$t\n";
    $sum += $t;

#   [...]

}

sub tau {
    my ($i, $j) = @_;
    my ($con, $dis, $tie_i, $tie_j, $n, @pairs);
    my (@runs, %runstat);

    for my $run (sort { $rank[$i]{$b} <=> $rank[$i]{$a} }
         keys %{ $rank[$i] }) {
    push @pairs, [ $rank[$i]{$run}, $rank[$j]{$run} ];
    push @runs, $run;
    }

#   [...]

    return $t;
}
```

The CLEF Initiative Portal

The CLEF Initiative (Conference and Labs of the Evaluation Forum) - CLEF2012 Working Notes

www.clef-initiative.eu/edition/clef2012/working-notes

Apple Yahoo! Google Maps YouTube Wikipedia Notizie I più conosciuti Università d...o di Padova

Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) - Main | The CLEF Initiative (Conference and Labs of the Evaluation Forum) - CLEF2012 Working Notes

The CLEF Initiative
Conference and Labs of the Evaluation Forum

Unlocking Information Access

Home | Editions | Tracks | Datasets | Publications | Links | About | Media Center | Contacts | Search

CLEF Initiative > Editions > CLEF2012 > Working Notes

CLEF2012

Agenda

Working Notes

CLEF 2012 Working Notes

Working notes for the CLEF 2012 Conference held in Rome, Italy.
Click here to show the working notes papers for all the tracks in the CLEF edition.

Contents

Cultural Heritage in CLEF (CHiC)

Cultural Heritage in CLEF (CHiC) Overview 2012
Vivien Petras, Nicola Ferro, Maria Gäde, Antoine Isaac, Michael Kleineberg, Ivano Masiero, Mattia Nicchio, and Juliane Stiller [paper]

The Sheffield and Basque Country Universities Entry to CHiC: Using Random Walks and Similarity to Access Cultural Heritage
Eneko Agirre, Paul Clough, Samuel Fernando, Mark Hall, Arantxa Otegi, and Mark Stevenson [paper]

UniNE at CLEF 2012
Mitra Akasereh, Nada Naji, and Jacques Savoy [paper]

Chemnitz at the CHiC Evaluation Lab 2012: Creating an Xtrieval Module for Semantic Enrichment
Jens Kürsten, Thomas Wilhelm, Daniel Richter, and Maximilian Eibl [paper]

Dealing with Sparse Document and Topic Representations: Lab Report for CHiC 2012
Philipp Schaer, Daniel Hienert, Frank Sawitzki, Andias Wira-Alam, and Thomas Lüke [paper]

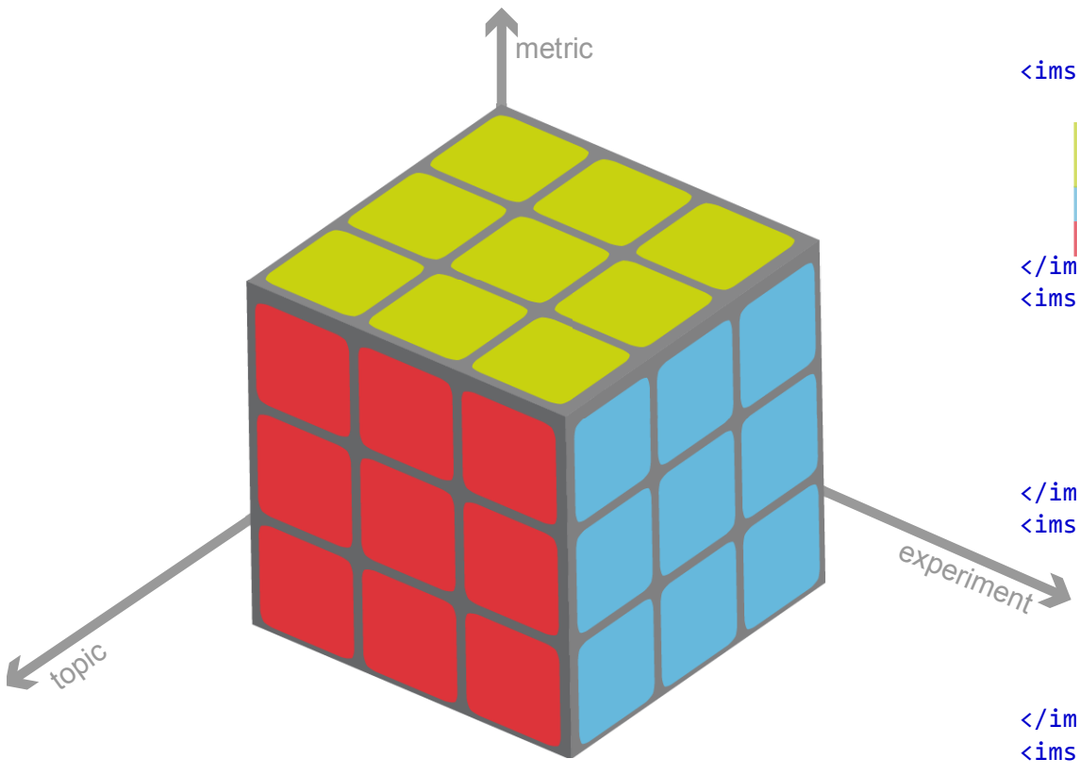
Query Expansion Using Wikipedia and Dbpedia
Nitish Aggarwal and Paul Buitelaar [paper]

Retrieval in the Intellectual Property Domain (CLEF-IP)

Image Retrieval in CLEF (ImageCLEF)

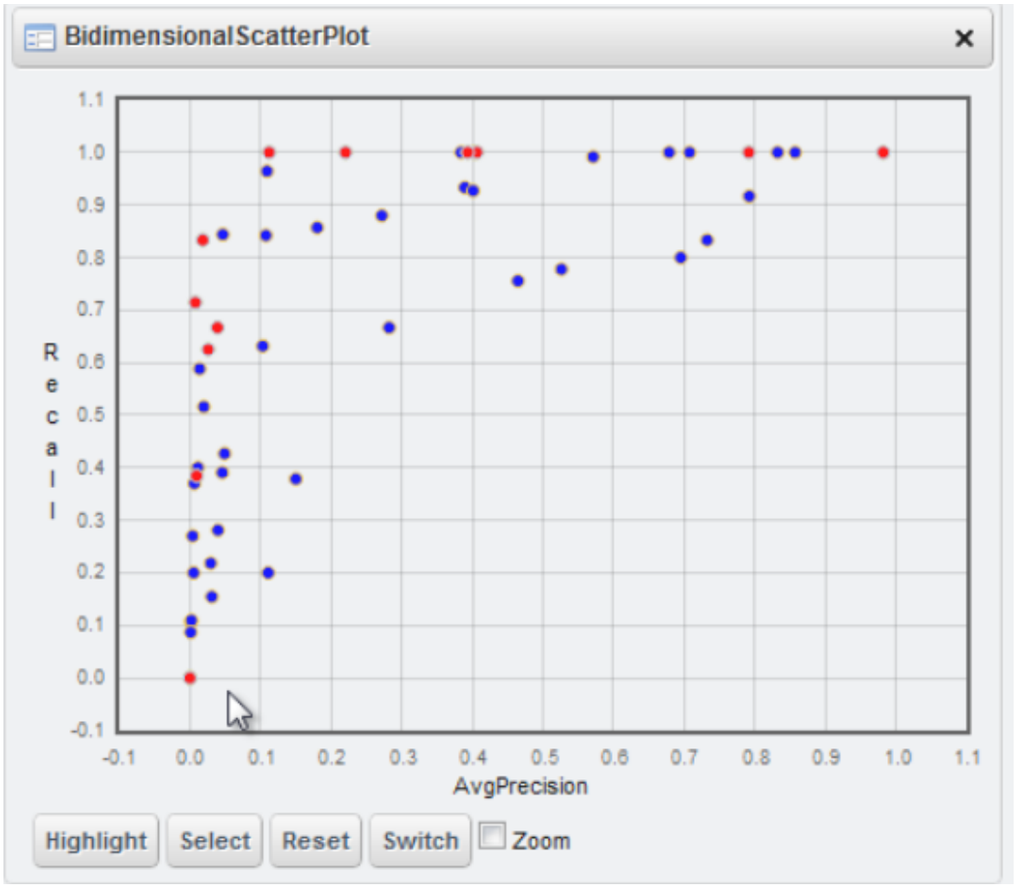
Visualizations

GET /task/CHIC-AH-MONO-EN-CLEF2012/measure




```
<ims:measure ims:identifier="177bcef2-00a0-4f59-b781-f285610f1c6f" ims:value="3.59523803E-1"
        ims:created="2012-10-11T16:23:31.011+02:00" ims:last-modified="2012-10-11T16:23:31.011+02:00">
    <ims:concept ims:identifier="AVERAGE_PRECISION"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-001" />
</ims:measure>
<ims:measure ims:identifier="00681943-2249-4a56-9c1e-5dbf82dcd1b5" ims:value="0.0E0"
    ims:created="2012-10-11T16:23:30.698+02:00" ims:last-modified="2012-10-11T16:23:30.698+02:00">
    <ims:concept ims:identifier="R_PRECISION"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-001" />
</ims:measure>
<ims:measure ims:identifier="0339409c-2baa-44e1-b977-48e4d79dc0ed" ims:value="5.0E-1"
    ims:created="2012-10-11T16:23:30.815+02:00" ims:last-modified="2012-10-11T16:23:30.815+02:00">
    <ims:concept ims:identifier="PRECISION_AT_0_INTERPOLATED_RECALL_LEVEL"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-001" />
</ims:measure>
<ims:measure ims:identifier="006df09d-157a-466c-bf2e-c0b3b10b6bc0" ims:value="4.4742262363E-1"
     ims:created="2012-10-11T16:23:30.882+02:00" ims:last-modified="2012-10-11T16:23:30.882+02:00">
    <ims:concept ims:identifier="AVERAGE_PRECISION"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-002" />
</ims:measure>
<ims:measure ims:identifier="164ef66a-aa48-413f-a229-540bb19f2d2e" ims:value="5.9090906382E-1"
    ims:created="2012-10-11T16:23:30.760+02:00" ims:last-modified="2012-10-11T16:23:30.760+02:00">
    <ims:concept ims:identifier="R_PRECISION"
         ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-002" />
</ims:measure>
<ims:measure ims:identifier="29a91449-61ae-49ef-96dc-2ae67377ef86" ims:value="1.0E0"
    ims:created="2012-10-11T16:23:30.594+02:00" ims:last-modified="2012-10-11T16:23:30.594+02:00">
    <ims:concept ims:identifier="PRECISION_AT_0_INTERPOLATED_RECALL_LEVEL"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-002" />
</ims:measure>
```

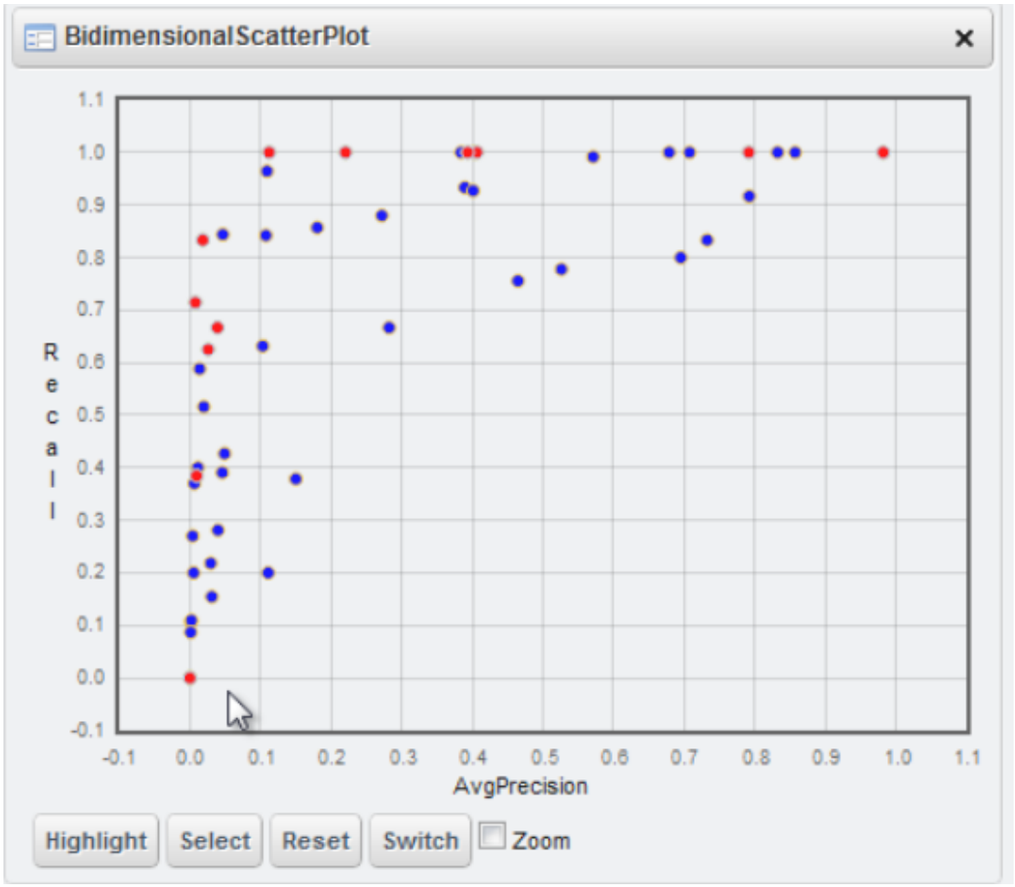
Visualizations




```
<ims:measure ims:identifier="177bcef2-00a0-4f59-b781-f285610f1c6f" ims:value="3.59523803E-1"
        ims:created="2012-10-11T16:23:31.011+02:00" ims:last-modified="2012-10-11T16:23:31.011+02:00">
    <ims:concept ims:identifier="AVERAGE_PRECISION"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-001" />
</ims:measure>
<ims:measure ims:identifier="00681943-2249-4a56-9c1e-5dbf82dcd1b5" ims:value="0.0E0"
    ims:created="2012-10-11T16:23:30.698+02:00" ims:last-modified="2012-10-11T16:23:30.698+02:00">
    <ims:concept ims:identifier="R_PRECISION"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-001" />
</ims:measure>
<ims:measure ims:identifier="0339409c-2baa-44e1-b977-48e4d79dc0ed" ims:value="5.0E-1"
    ims:created="2012-10-11T16:23:30.815+02:00" ims:last-modified="2012-10-11T16:23:30.815+02:00">
    <ims:concept ims:identifier="PRECISION_AT_0_INTERPOLATED_RECALL_LEVEL"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-001" />
</ims:measure>
<ims:measure ims:identifier="006df09d-157a-466c-bf2e-c0b3b10b6bc0" ims:value="4.4742262363E-1"
     ims:created="2012-10-11T16:23:30.882+02:00" ims:last-modified="2012-10-11T16:23:30.882+02:00">
    <ims:concept ims:identifier="AVERAGE_PRECISION"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-002" />
</ims:measure>
<ims:measure ims:identifier="164ef66a-aa48-413f-a229-540bb19f2d2e" ims:value="5.9090906382E-1"
    ims:created="2012-10-11T16:23:30.760+02:00" ims:last-modified="2012-10-11T16:23:30.760+02:00">
    <ims:concept ims:identifier="R_PRECISION"
         ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-002" />
</ims:measure>
<ims:measure ims:identifier="29a91449-61ae-49ef-96dc-2ae67377ef86" ims:value="1.0E0"
    ims:created="2012-10-11T16:23:30.594+02:00" ims:last-modified="2012-10-11T16:23:30.594+02:00">
    <ims:concept ims:identifier="PRECISION_AT_0_INTERPOLATED_RECALL_LEVEL"
        ims:namespace="http://direct.dei.unipd.it/metrics/" />
    <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
    <ims:topic ims:identifier="CHIC-002" />
</ims:measure>
```

Visualizations

POST /visualization/




```
<ims:visualization ims:identifier="3dedf5b5-5026-4054-9fc6-827512791960" ims:scope="PUBLIC">
    <ims:concept ims:identifier="SCATTERPLOT"
        ims:namespace="http://direct.dei.unipd.it/visualizations/" />
    <ims:parameters>
        <ims:parameter>
            <ims:concept ims:identifier="X_AXIS"
                ims:namespace="http://direct.dei.unipd.it/visualizations/" />
            <ims:value>
                <ims:concept ims:identifier="AVERAGE_PRECISION"
                    ims:namespace="http://direct.dei.unipd.it/metrics/" />
            </ims:value>
        </ims:parameter>
        <ims:parameter>
            <ims:concept ims:identifier="Y_AXIS"
                ims:namespace="http://direct.dei.unipd.it/visualizations/" />
            <ims:value>
                <ims:concept ims:identifier="RECALL"
                    ims:namespace="http://direct.dei.unipd.it/metrics/" />
            </ims:value>
        </ims:parameter>
        <ims:parameter>
            <ims:concept ims:identifier="HIGHLIGHT"
                ims:namespace="http://direct.dei.unipd.it/visualizations/" />
            <ims:value>
                <ims:measure ims:identifier="177bcef2-00a0-4f59-b781-f285610f1c6f"/>
                <ims:measure ims:identifier="006df09d-157a-466c-bf2e-c0b3b10b6bc0"/>
            </ims:value>
        </ims:parameter>
    </ims:parameters>
    <ims:measures>
        <ims:measure ims:identifier="177bcef2-00a0-4f59-b781-f285610f1c6f" ims:value="3.59523803E-1" >
            <ims:concept ims:identifier="AVERAGE_PRECISION"
                ims:namespace="http://direct.dei.unipd.it/metrics/" />
            <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
                <ims:topic ims:identifier="CHIC-001" />
        </ims:measure>
        <ims:measure ims:identifier="006df09d-157a-466c-bf2e-c0b3b10b6bc0" ims:value="4.4742262363E-1">
            <ims:concept ims:identifier="AVERAGE_PRECISION"
                ims:namespace="http://direct.dei.unipd.it/metrics/" />
            <ims:experiment ims:identifier="CHIC-AH-MONO-EN-CLEF2012.ARANTZA.OTEGI@EHU.ES.EXP_UKB_WN100" />
            <ims:topic ims:identifier="CHIC-002" />
        </ims:measure>
    </ims:measures>
    <ims:snapshots>
        <ims:snapshot ims:identifier="017c333a-4b7c-4267-926d-f15fe3554efd" ims:media-type="application/pdf" >
            <ims:content ims:content-transfer-encoding="base64">
                c2lnbiBjb250ZW50
            </ims:content>
        </ims:snapshot>
    </ims:snapshots>
</ims:visualization>
```

Actionable Papers

CHIC2012-Overview-Paper-updated-2012-09-03.htm

Cultural Heritage in CLEF (CHiC) Overview 2012

Vivien Petras, Nicola Ferro, Maria Gäde, Antoine Isaac, Michael Kleineberg, Ivano Masiero, Mattia Nicchio and Juliane Stiller

1 Introduction

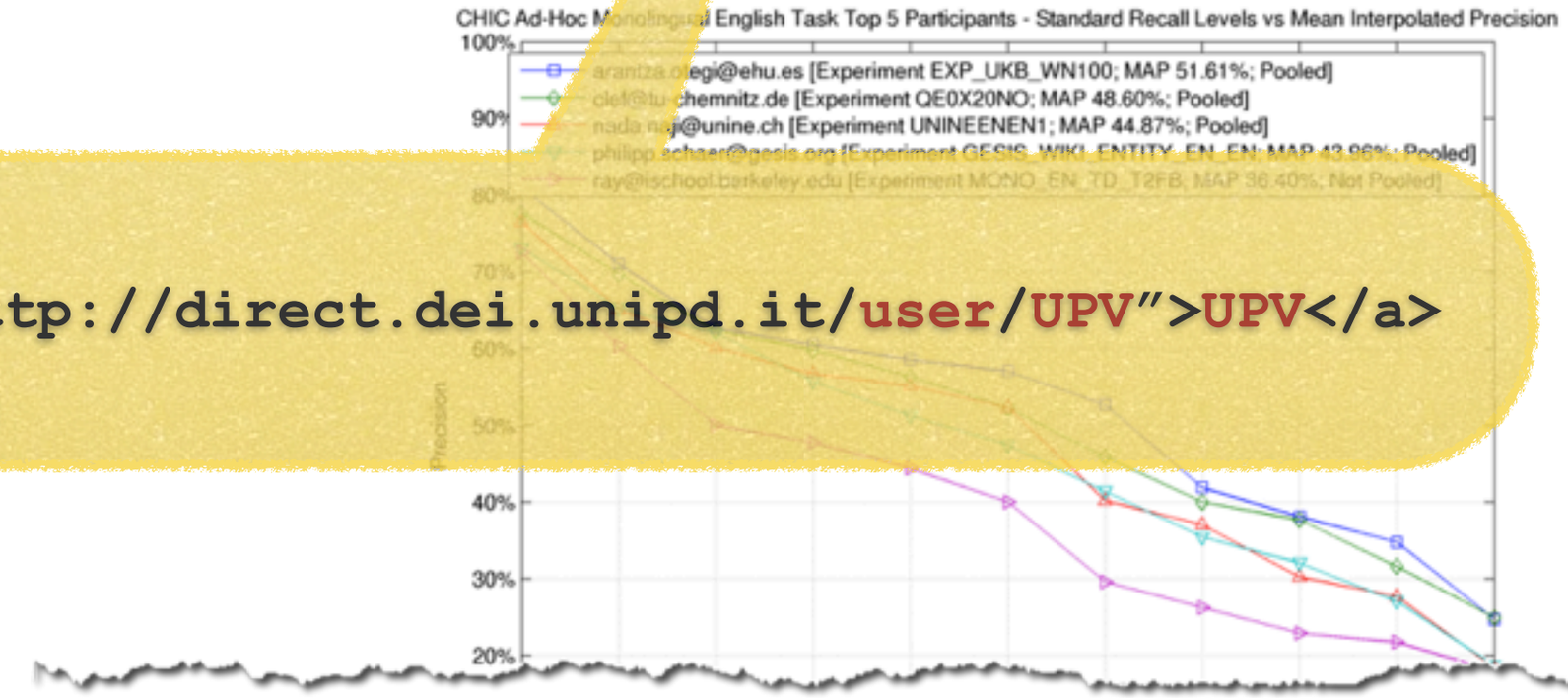
Cultural heritage content is often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of different levels of complexity....

2 Results Analysis

Table 1. Best monolingual Experiments and Performance Difference between best and last (up to 5) Experiment (in MAP)

Track	Rank	Part.	Experiment Identifier	MAP
Monolingual English	1st	UPV	EXP_UKB_WN100	51.61%
	2nd	Chemnitz	QE0X20NO	48.60%
	3rd	Neuchatel	UNINEENEN1	44.87%
	4th	Gesis	GESIS_WIKI_ENTITY_EN_EN	43.96%
	5th	Berkeley	MONO_EN_TD_T2FB	36.40%
	Diff.			41.78%

Figures 7 to 9 show the interpolated recall vs. average precision for the top groups of the monolingual tasks.




```
<a href="http://direct.dei.unipd.it/user/UPV">UPV</a>
```

Actionable Papers

CHIC2012-Overview-Paper-updated-2012-09-03.htm

Cultural Heritage in CLEF (CHiC) Overview 2012

Vivien Petras, Nicola Ferro, Maria Gäde, Antoine Isaac, Michael Kleineberg, Ivano Masiero, Mattia Nicchio and Juliane Stiller

1 Introduction

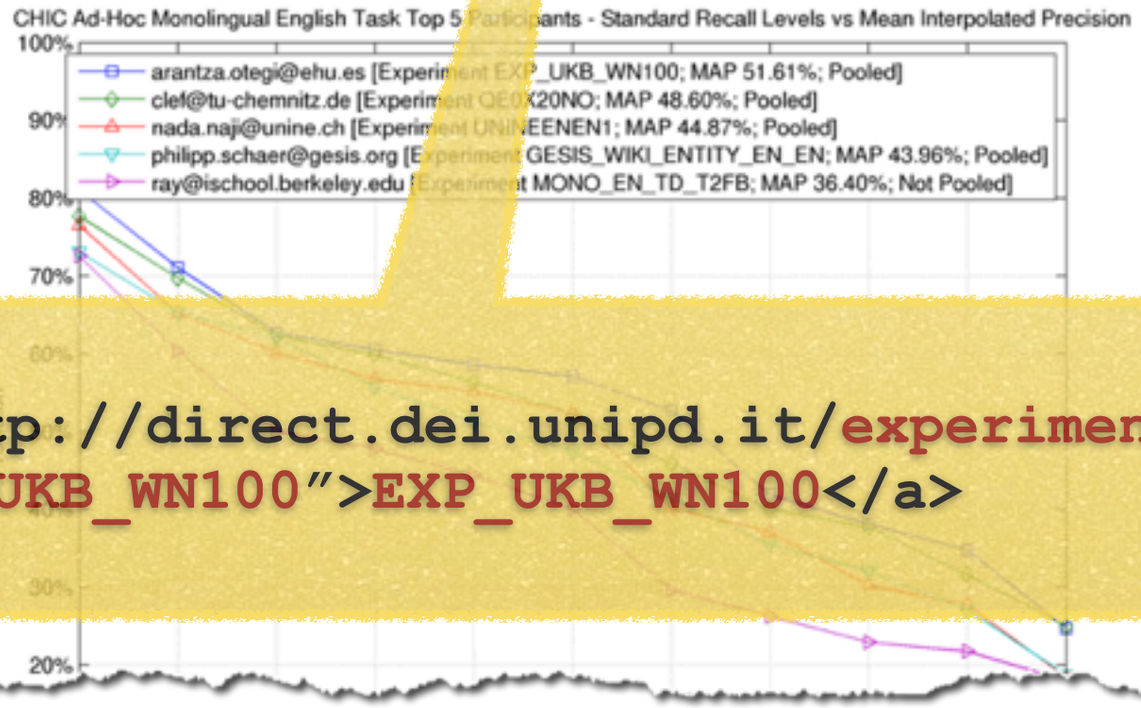
Cultural heritage content is often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of different levels of complexity....

2 Results Analysis

Table 1. Best monolingual Experiments and Performance Difference between best and last (up to 5) Experiment (in MAP)

Track	Rank	Part.	Experiment Identifier	MAP
Monolingual English	1st	UPV	EXP_UKB_WN100	51.61%
	2nd	Chemnitz	QE0X20NO	48.60%
	3rd	Neuchatel	UNINEENEN1	44.87%
	4th	Gesis	GESIS_WIKI_ENTITY_EN_EN	43.96%
	5th	Berkeley	MONO_EN_TD_T2FB	36.40%
	Diff.			41.78%

Figures 7 to 9 show the interpolated recall vs. average precision for the top groups of the monolingual tasks.




```
<a href="http://direct.dei.unipd.it/experiment/
EXP_UKB_WN100">EXP_UKB_WN100</a>
```

Actionable Papers

CHIC2012-Overview-Paper-updated-2012-09-03.htm

Cultural Heritage in CLEF (CHiC) Overview 2012

Vivien Petras, Nicola Ferro, Maria Gäde, Antoine Isaac, Michael Kleineberg, Ivano Masiero, Mattia Nicchio and Juliane Stiller

1 Introduction

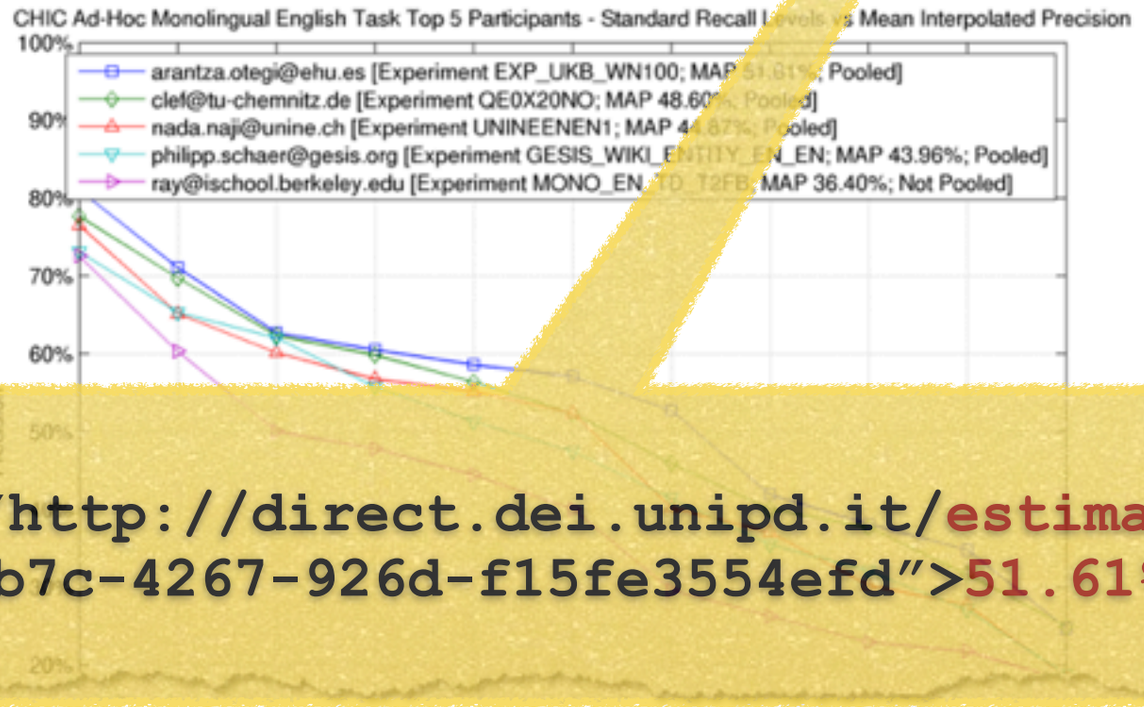
Cultural heritage content is often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of different levels of complexity....

2 Results Analysis

Table 1. Best monolingual Experiments and Performance Difference between best and last (up to 5) Experiment (in MAP)

Track	Rank	Part.	Experiment Identifier	MAP
Monolingual English	1st	UPV	EXP_UKB_WN100	51.61%
	2nd	Chemnitz	QE0X20NO	48.60%
	3rd	Neuchatel	UNINEENEN1	44.87%
	4th	Gesis	GESIS_WIKI_ENTITY_EN_EN	43.96%
	5th	Berkeley	MONO_EN_TD_T2FB	36.40%
	Diff.			41.78%

Figures 7 to 9 show the interpolated recall vs. average precision for the top groups of the monolingual tasks.




```
<a href="http://direct.dei.unipd.it/estimate/
017c333a-4b7c-4267-926d-f15fe3554efd">51.61%</a>
```

Actionable Papers

```
<img src="http://direct.dei.unipd.it/visualization/
017c333a-4b7c-4267-926d-f15fe3554efd/snapshot/
177bcef2-00a0-4f59-b781-f285610f1c6f"/>
```

CHiC2012-Overview-Paper-Updated-2012-08-03.tex

Cultural Heritage in CLEF (CHiC) Overview 2012

and Juliane Stiller

Introduction

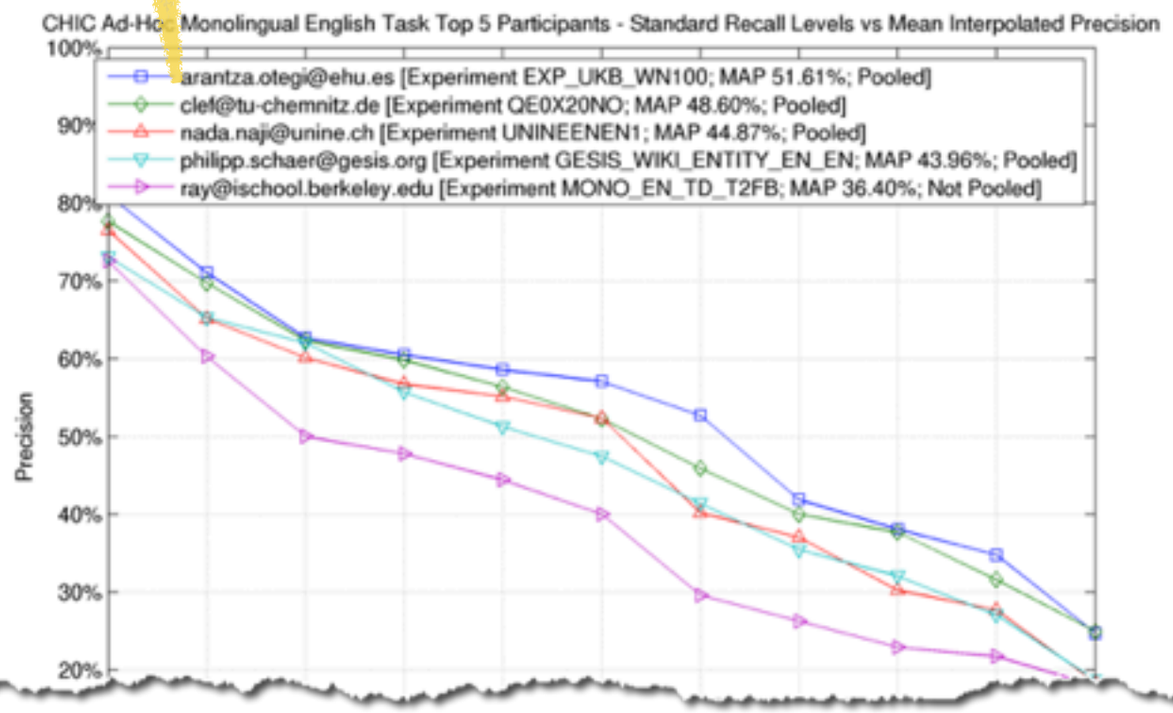
Cultural heritage content is often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of different levels of complexity....

2 Results Analysis

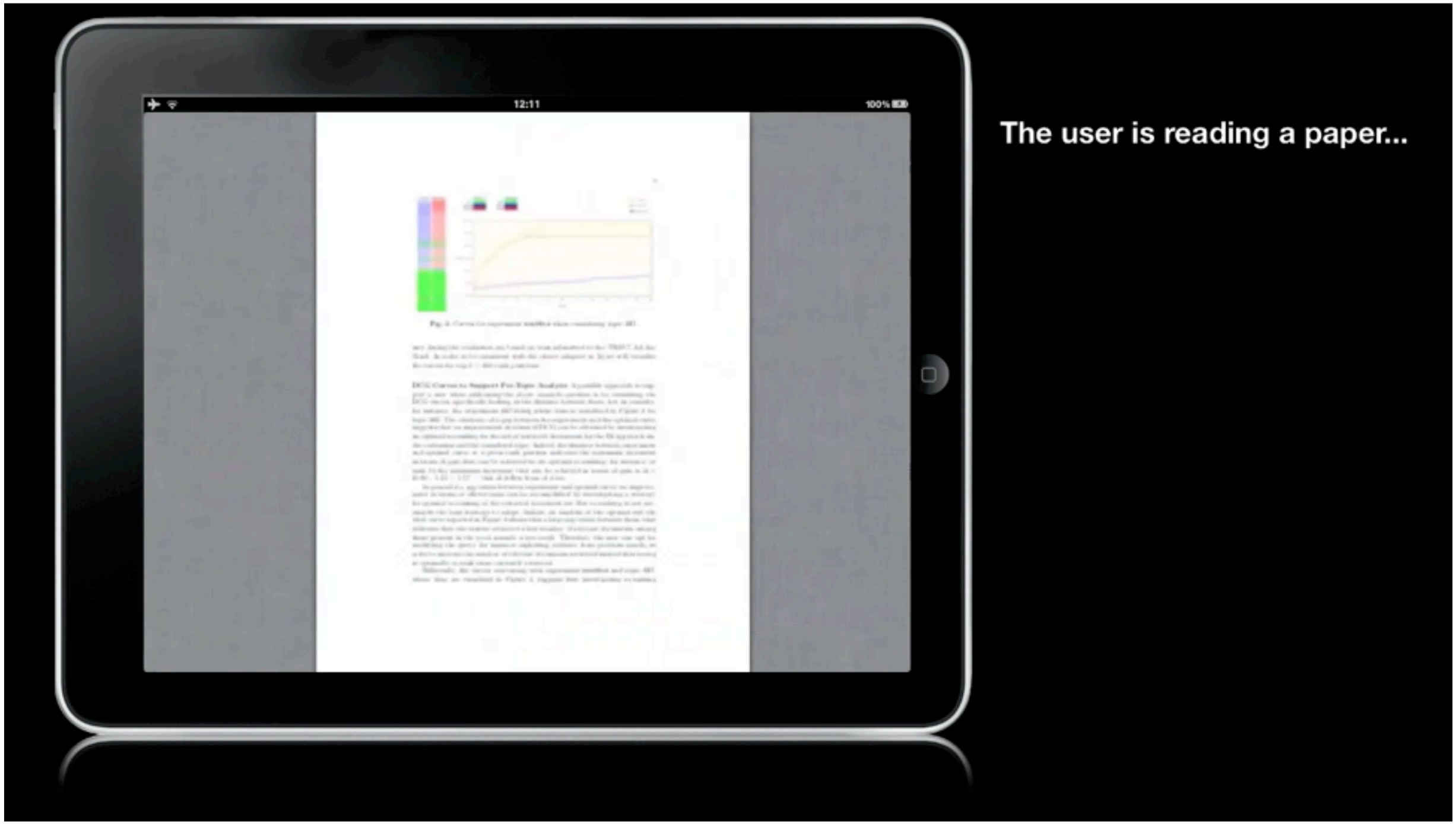
Table 1. Best monolingual Experiments and Performance Difference between best and last (up to 5) Experiment (in MAP)

Track	Rank	Part.	Experiment Identifier	MAP
Monolingual English	1st	UPV	EXP_UKB_WN100	51.61%
	2nd	Chemnitz	QE0X20NO	48.60%
	3rd	Neuchatel	UNINEENEN1	44.87%
	4th	Gesis	GESIS_WIKI_ENTITY_EN_EN	43.96%
	5th	Berkeley	MONO_EN_TD_T2FB	36.40%
	Diff.			41.78%

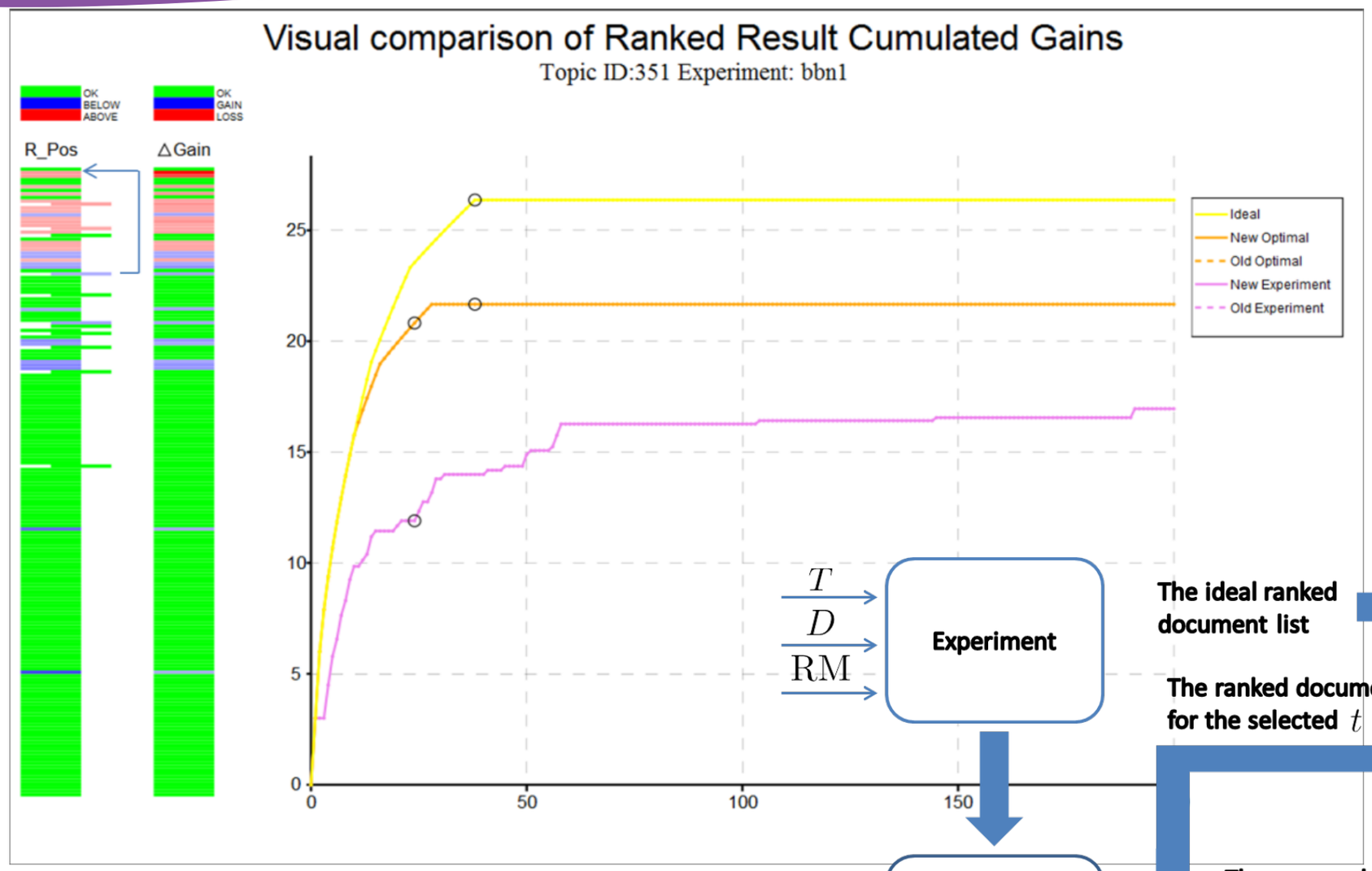
Figures 7 to 9 show the interpolated recall vs. average precision for the top groups of the monolingual tasks.

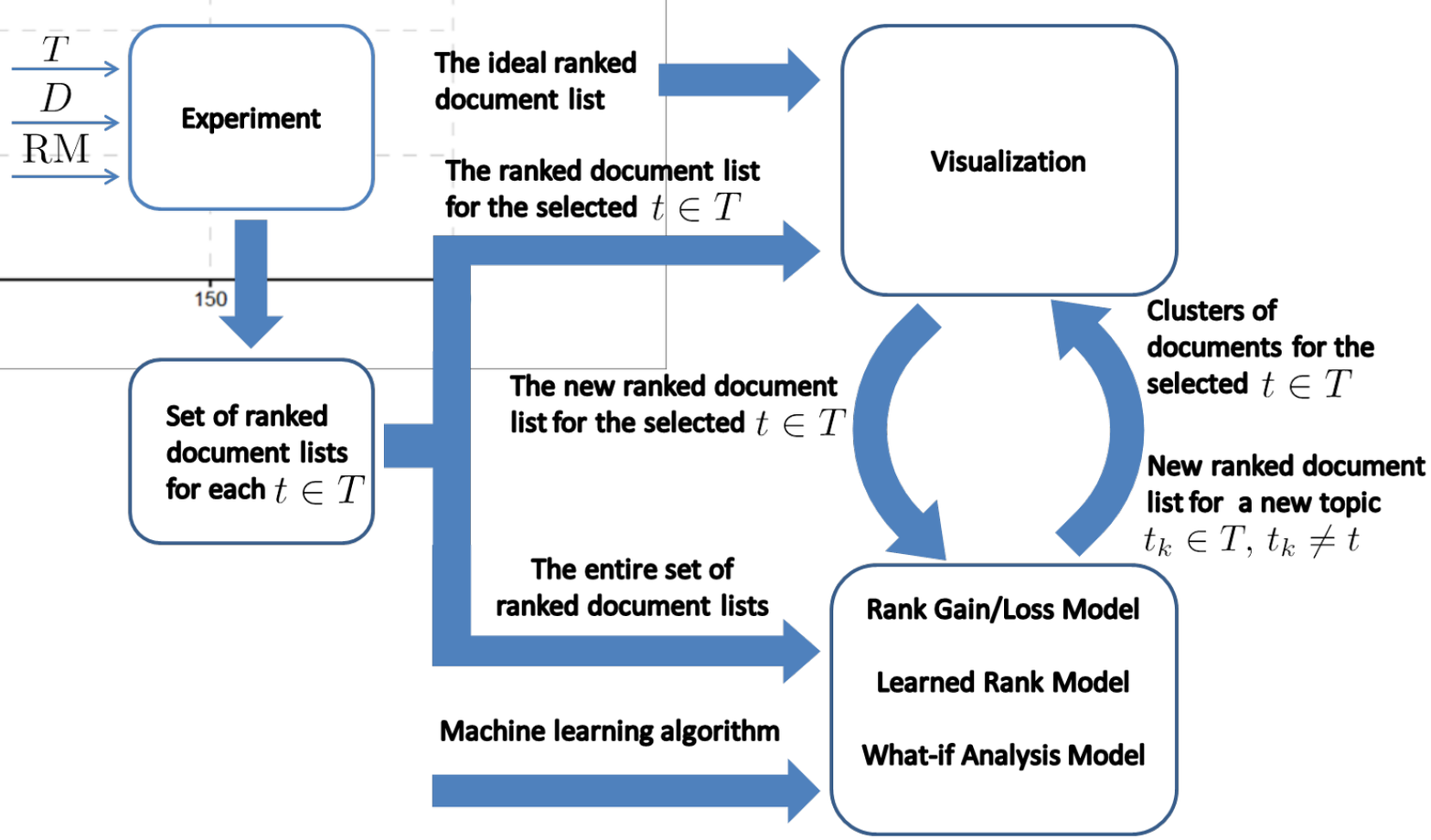


Actionable Papers and Visual Analytics

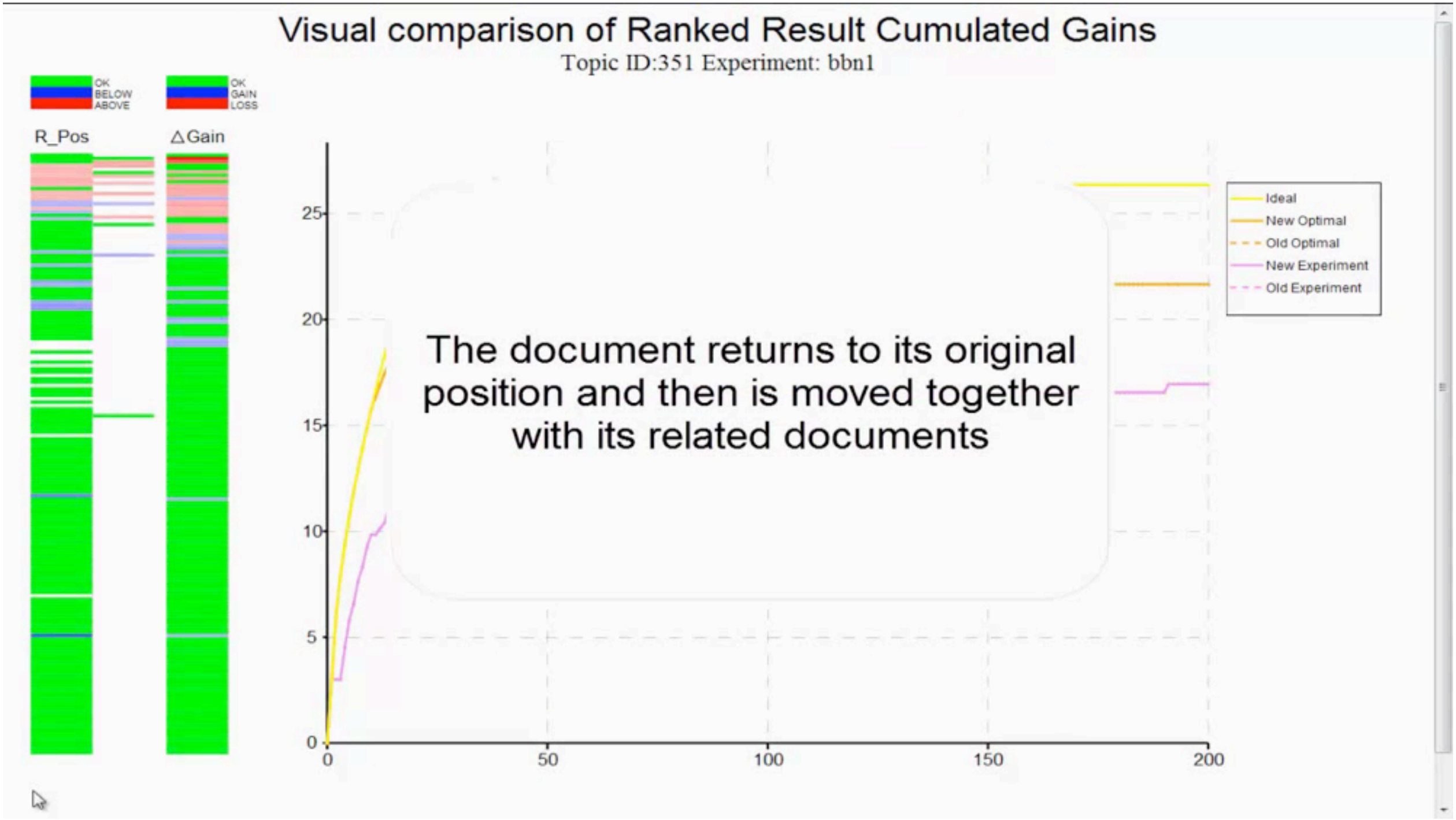


Visual What-if Failure Analysis

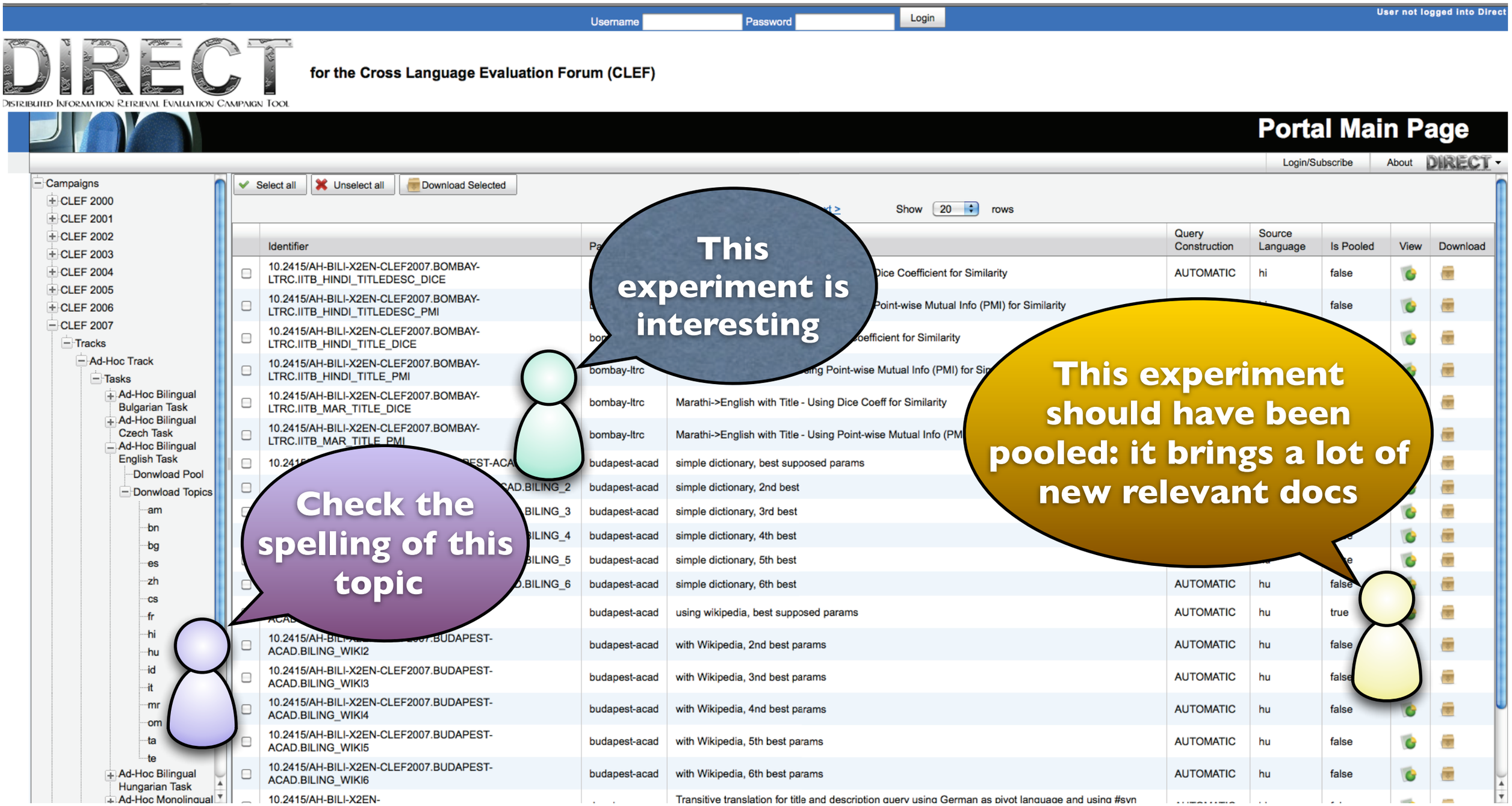




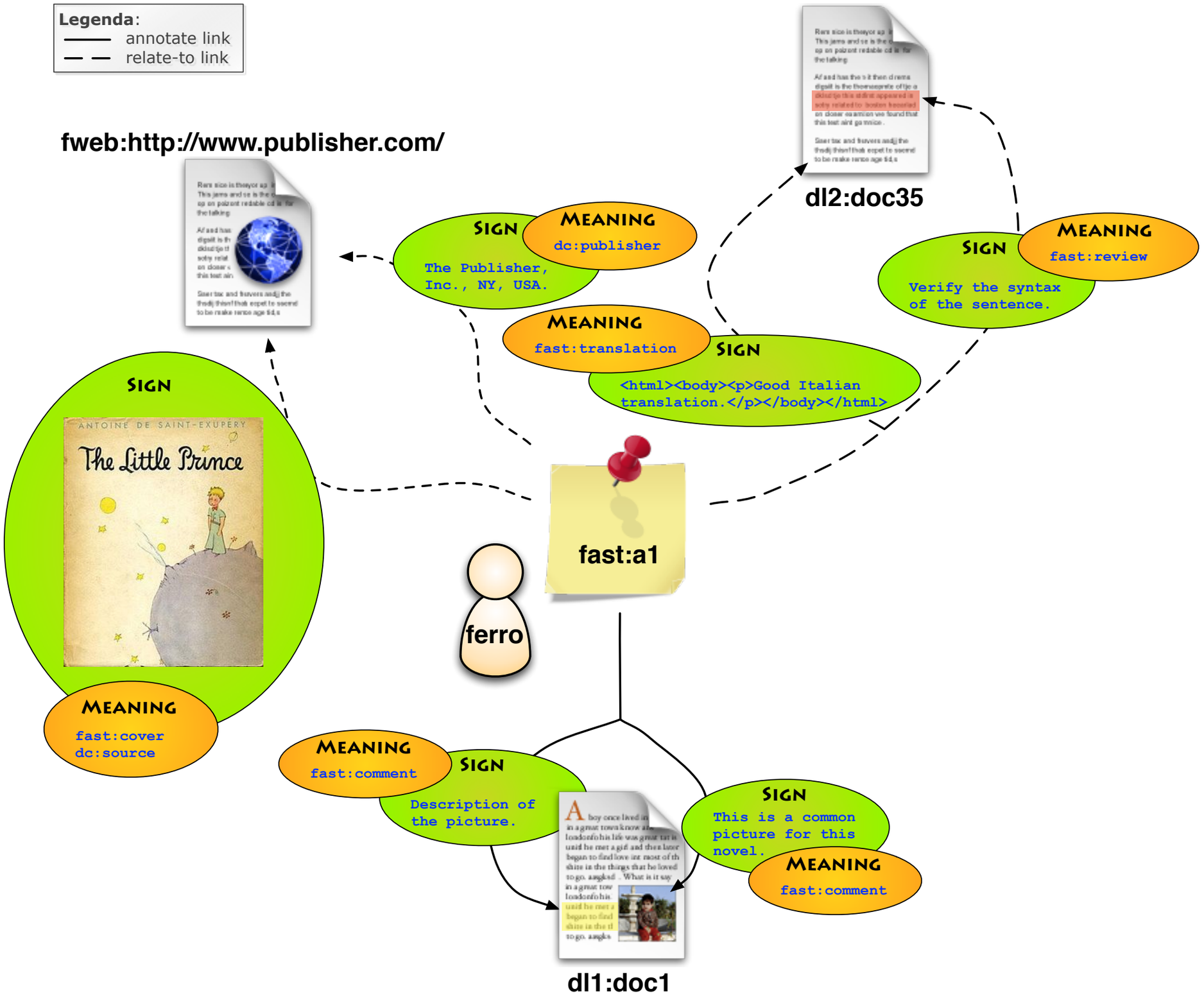
Visual What-if Failure Analysis



User Annotations



User Annotations



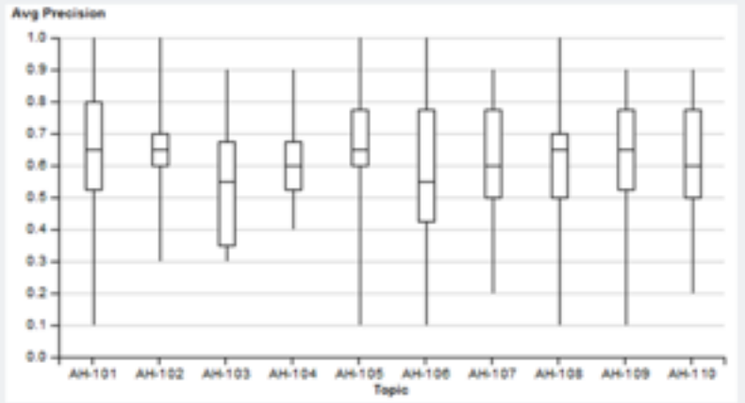
User Annotations

TableChart

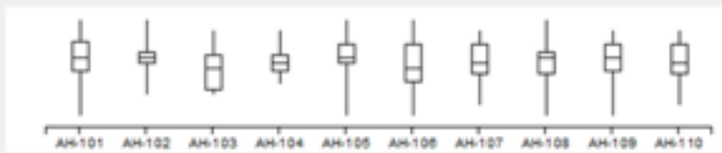
Average Precision

topic	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10
AH-101	0.2	0.8	0.6	0.5	0.8	0.1	0.6	0.7	1.0	0.9
AH-102	0.3	0.6	0.7	0.6	0.7	0.4	0.7	0.6	0.9	1.0
AH-103	0.3	0.5	0.9	0.3	0.5	0.3	0.6	0.6	0.8	0.7
AH-104	0.5	0.6	0.6	0.7	0.6	0.4	0.9	0.5	0.6	0.7
AH-105	0.1	0.7	1.0	0.7	0.6	0.3	0.8	0.9	0.6	0.6
AH-106	0.1	0.4	1.0	0.6	0.7	0.2	0.5	1.0	0.5	0.8
AH-107	0.4	0.8	0.9	0.7	0.5	0.2	0.5	0.7	0.9	0.5
AH-108	0.3	1.0	0.7	0.5	0.7	0.1	0.7	0.5	0.8	0.6
AH-109	0.2	0.8	0.8	0.9	0.6	0.1	0.6	0.7	0.7	0.5
AH-110	0.2	0.6	0.7	0.6	0.8	0.5	0.5	0.5	0.8	0.9

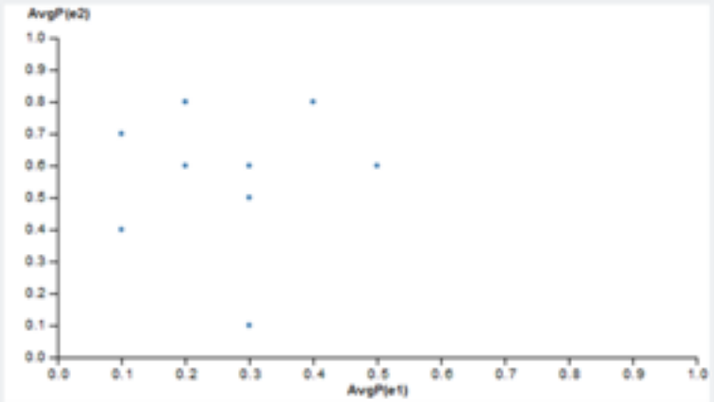
BoxPlotChart



Overview Grouping ReferenceLine Settings Ordering

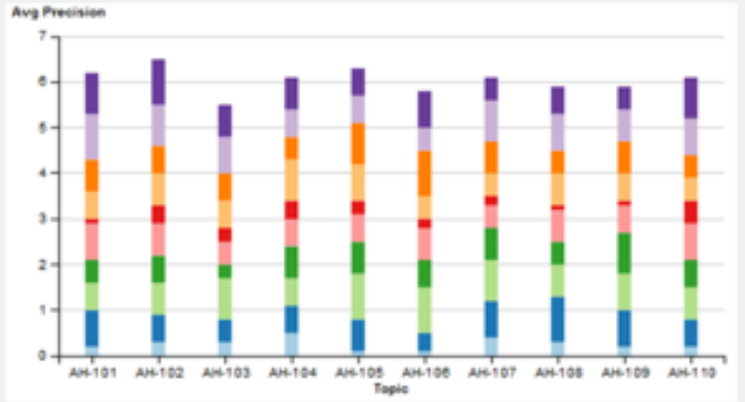


ScatterPlotChart



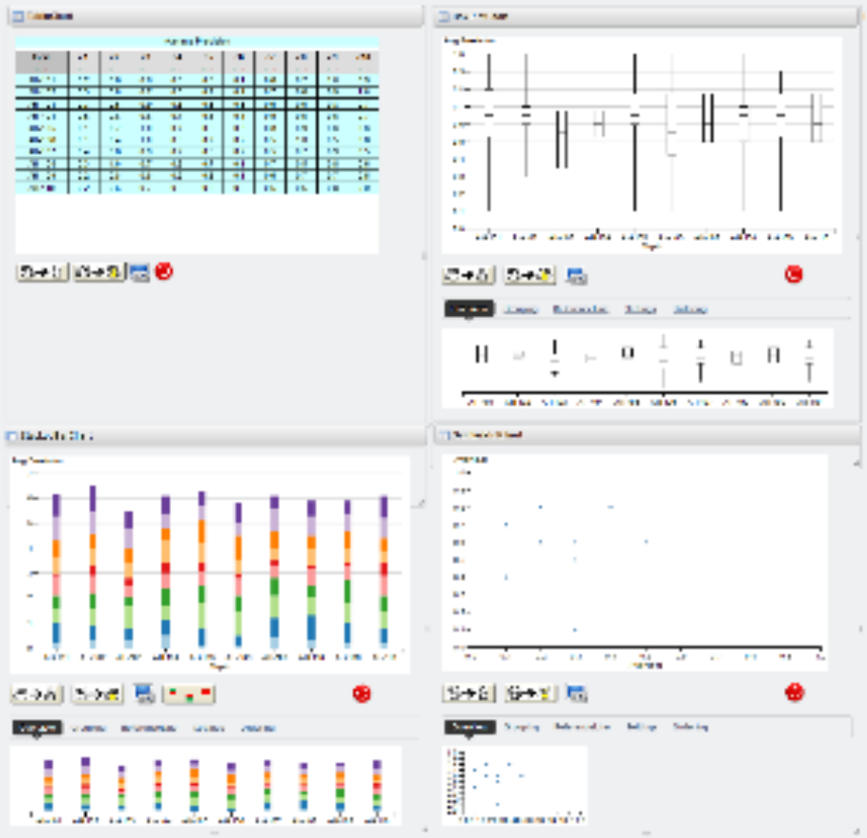
Overview Grouping ReferenceLine Settings Ordering

StackedBarChart



Overview Grouping ReferenceLine Settings Ordering

User Annotations



- good results for experiments on topic 351

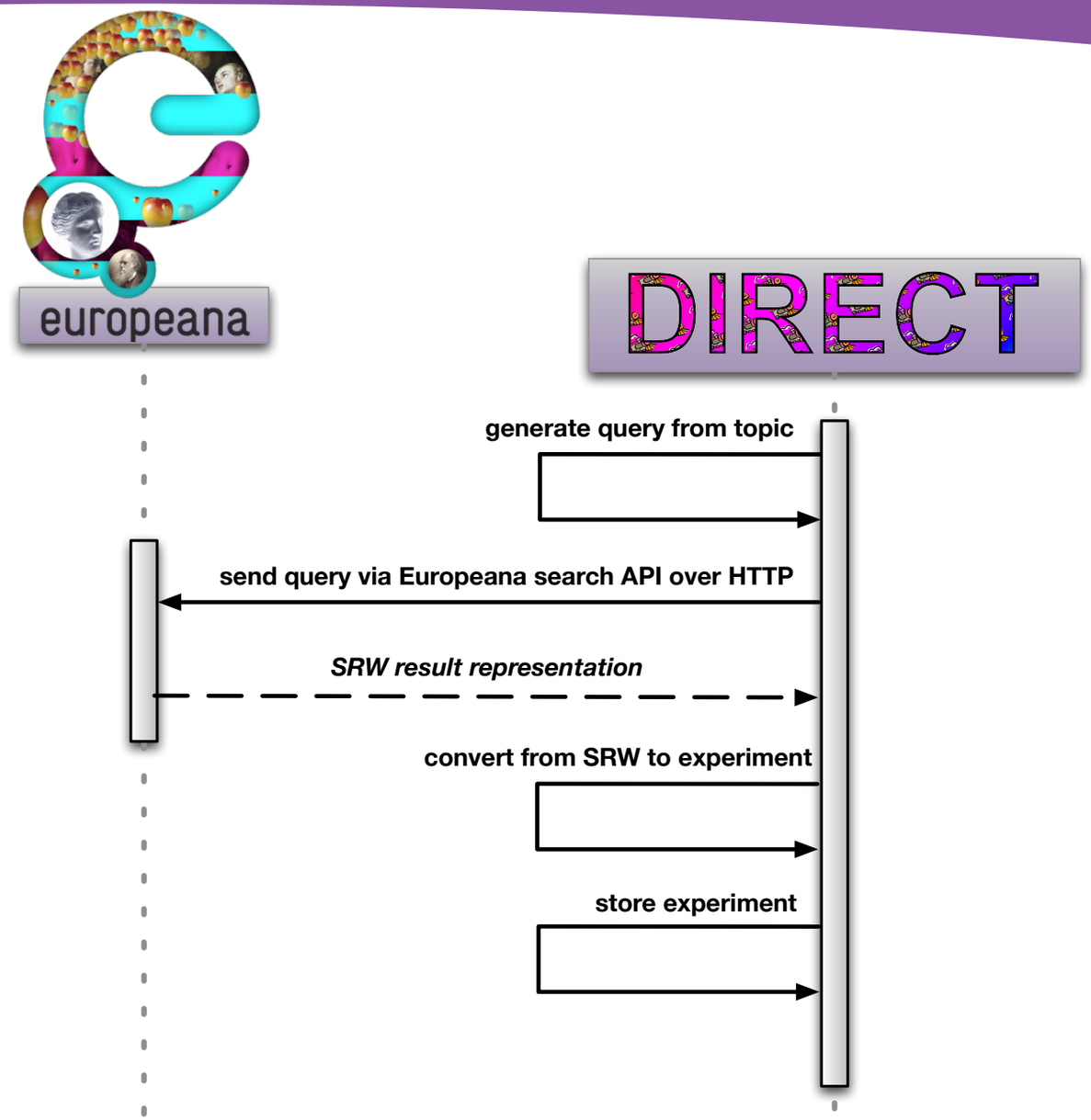
- please check correlation trend on scatter plot

☑ Table chart ☐ BoxPlot chart ☐ Stacked Bar chart ☑ Scatter Plot chart

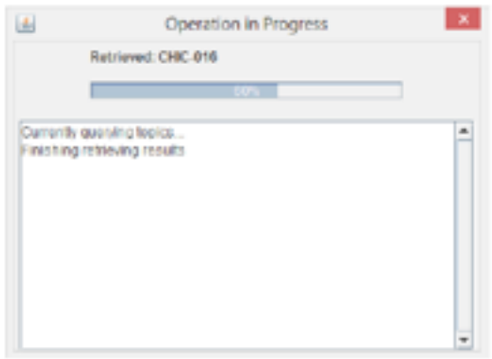
SUBMIT

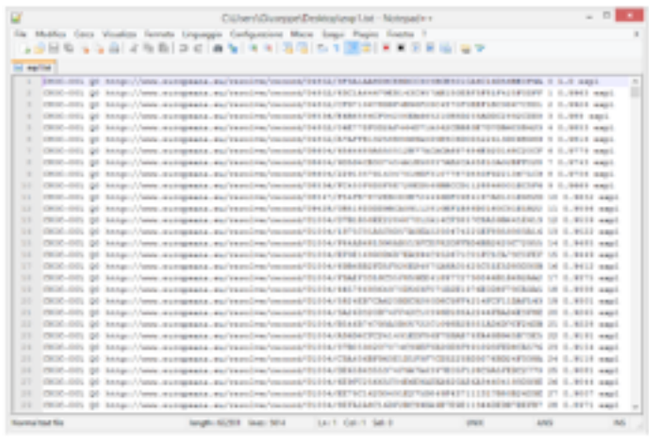
Continuous Evaluation

- Programmatically evaluate a system under examination
  - somewhat similar to unit testing in software engineering



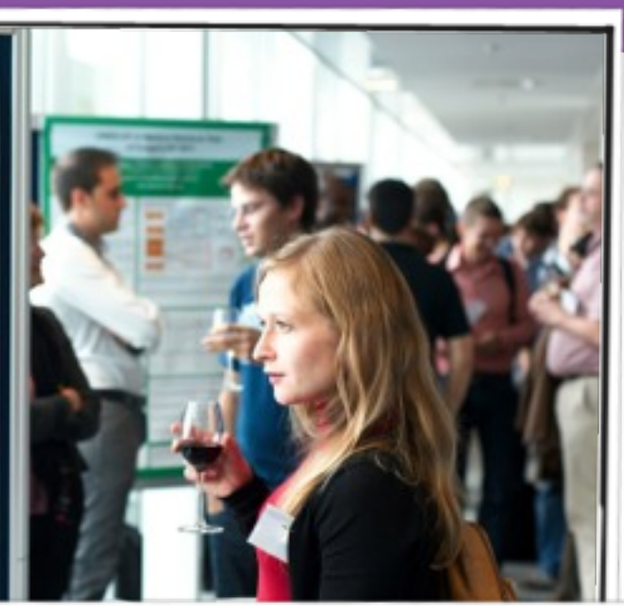
- Currently experimented with Europeana
  - based on the CHiC test collections





Sharing and Extending

The Community



Final Remarks

- We have discussed why and how to design an evalution infrastructure
  - bridging between IR and DB is a key factor

- The need for an evaluation infrastructure is raising in the community
  - DESIRE 2011, SWIRL II, PROMISE Retreat
  - EvaluitIR, Apache Open Relevance Project

- To be successful, enablers are needed
  - community adoption, modification, and extension
  - “political” willingness
  - economical sustainability

Thank You