Authenticus - Enabling the Identification and Validation of Portuguese Publications

Fernando Silva (fds@dcc.fc.up.pt)

14th November 2013

Collaborators:

Sylvia Bugla, Fábio Domingues, João Rodrigues

http://authenticus.up.pt
Authenticus – Goals

- Develop a validated database of Portuguese publications
  - i.e. authored by researchers associated with Portuguese institutions.
- Automatically associate publications to researchers and institutions
  - Author names and addresses disambiguation
  - Difficult problem
- Allow user validation with minimal disruption
- Allow interoperability of data
  - Import/Export of publications for researchers and institutions
  - Import publications data from ISI, Scopus, DBLP
  - Export publications to other systems: SIGARRA, DeGóis, etc.
Authenticus – what for?

• Avoid manually insertion of publications by researchers
• More reliable bibliometric indicators on individuals
  • to assist peer-reviewing
  • more trustable metadata and indicators
  • individual research assessments, e.g. promotion contests
  • propagate data to institutions
• Bibliometric indicators within Universities
  • assist internal assessments of faculty performance
  • indicators for management purposes
• FCT – researchers and research units performance assessments
• Others …
1. Database of researchers and Institutions
   • source: FCT, REBIDES, Institutions (UP), MCTES (MEC), Algorithm

2. Database of publications
   • source: ISI, Scopus, DBLP, etc.

3. Author name and address identification algorithm
   • Multi-criteria classification algorithm

4. Researchers interface
   • Engage researchers in confirming their publications
   • Allow export of publications lists

5. Implement bibliometric indicators for individual researchers
Authenticus: data import workflow

1. Import Publications
2. Redundancy checking
3. UPDATE of meta-data information
   - SciVerse
   - Article Match Retrieval

Database
Authenticus: workflow

Import Publications → Redundancy checking

UPDATE of meta-data information

Database

Data Sharing / Export

Web services, linked-data

Bibtex, ISI Exprt
RIS, EndNote,
MODS ...

Interface

Identification Alg.
Authenticus: data export

Tool
Bibutils: converts between various bibliography formats using a common intermediate format (MODS)

MODS
Metadata Object Description Schema
XML schema for a bibliographic element set used for library applications. (http://www.loc.gov/standards/mods/)

Data Export
Bibtex, EndNote, RIS, ISI, MODS XML, Word 2007 bib, ADS
Currently, we support two main service requests:

**GetUpdatedPublications**(ResearcherInternalID, PageSize, PageTimeStamp)

- requests for periodic synchronization of validated publications
- **ResearcherInternalID** - 1 user or ALL users from the requesting institution
- **PageSize** – specifies #publications to be returned in each response
- **PageTimeStamp** – the time at which last results were received

**GetResearchersWithPublicationsToConfirm**(PageTimeStamp)

- returns the researchers that have publications to be confirmed.
Request example:

```xml
<?xml version="1.0" encoding="UTF-8"?>
<authenticus>
  <request>
    <api_key>123456</api_key>
    <action>
      <name>getUpdatedPublications</name>
      <options>
        <filter>
          <researcher_internal_id>ALL</researcher_internal_id>
          <page_timestamp>2013-09-30 15:45:18.904389</page_timestamp>
        </filter>
        <get>fields_of_science</get>
        <get>subject_categories</get>
        <get>isi_info</get>
        <get>scopus_info</get>
        <get>journal_id</get>
        <get>researcher_internal_id</get>
        <page_size>50</page_size>
      </options>
    </action>
  </request>
</authenticus>
```

- Identifier of requesting institution
- Optional fields to be included in the web service response!
Response example:

```xml
<?xml version="1.0" encoding="UTF-8"?>
<authenticus>
  <request-info>
    <api_key>123456</api_key>
    <action>
      <name>getUpdatedPublications</name>
      <options>
        <filter>
          <page_timestamp>2013-09-30 15:45:18.904389</page_timestamp>
        </filter>
        <page_size>50</page_size>
      </options>
    </action>
  </request-info>
  <success> // warning message or error message
    <message>Ok</message>
    <results>
      <returned>50</returned>
      <remaining>11108</remaining>
      <next_page_timestamp>2013-11-01 15:43:05.000982</next_page_timestamp>
    </results>
  </success>
</authenticus>
```
Authenticus architecture

Three-layered architecture comprised of presentation, business, and data layers. Parallel cross-cutting layer with support modules. Name identification algorithm which communicates with the database.
Software:
SO: Linux Ubuntu server 12.04.02 LTS
PHP: 5.3.10-1ubuntu3.6 with Suhosin-Patch (cli)

Database Server:
MySQL Community Server (GPL) - Version: 5.6.11-log

Web Server:
HTTP Server: Apache 2.2.22
Database client version: libmysql - mysqli
PHP extension: mysqli

Web Interface:
Framework: CakePHP 2.3.4
Layout: Bootstrap 2.3.2
Javascript Libraries: jQuery v@1.8.0, jQuery UI - v1.8.23, Google Charts

Other Libraries:
python-mechanize, php5-curl, php5-mysqli, bibutils
Authenticus: identification algorithm

Finding Potential Matches

Filter: Name construction rules

ID  Author Name
---  ----------------------------------------
3435 Abel Dias dos Santos
1208 Alberto Carlos Pires Dias
7085 Alberto Romão Dias
3212 Alexandra da Costa Dias
... ...

Verification of Potential Matches

Verification Based on Name
Verification Based on Co-authors
Verification Based on Institutions
Verification Based on Keywords
Verification Based on Scientific Areas
Verification Based on Journals/Confs.

Decision

Database

Set of Potential Researchers For Analysed Author

Author Name Ex: (Dias, M.A)
Researchers: indicators

Indicators for a 3 to 5 year period.

Descriptive:
1. # of documents in journals Q1
2. Top 10%: # of documents within the 10% most cited
3. H-index (Hirsch, 2005)
4. # of documents fractioned
5. # of documents in journals Q1 and fractioned
6. Top 10%f – # of documents among the 10% more cited and fractioned

Quality:
1. % of documents in journals of Q1
2. Average of the normalized impact factor
3. Aggregated normalized and fractioned impact factor
4. Average impact normalized
5. % of documents among the 10% most cited
6. % of documents not cited
Authenticus: numbers

# Publications ISI: 163,609 (up to July 2013)

Researchers

# researchers PT (2007-2012): 41,309
# initial set of researchers in Authenticus: 77,430
# researchers with FCT_ID: 49,231
# researchers with UP_ID: 4,301
# researchers with REBIDES (2007-2012): 46,679

Institutions (including past institutions):

Universities/Polytechnics + all schools: 669
   → public: 533
   → private: 136

Main public universities: 16 / polytechnics: 18

Research Centers: 549
Associate Laboratories: 29
Authenticus: http://authenticus.up.pt

Live Demonstration

Fernando Manuel Augusto da Silva
Username: fmsilva@fc.up.pt  Group: Researcher  Federated User
Last Login: 2013-11-13 00:40:46
Email: fmsilva@fc.up.pt
Full Name: Fernando Manuel Augusto da Silva
Shibboleth Institution: Faculdade de Ciências da Universidade do Porto
Shibboleth Organisation: Universidade do Porto

Associated With: Fernando Manuel Augusto Silva
Future Work / Partnerships

University of Porto
• November 2013 – conclude test phase of Authenticus-SIGARRA communication
• December 2013 – hopefully all faculty will be using Authenticus
• November 2013 – identify Scopus and DBLP integration

DeGóis system:
• We are receptive to supply confirmed data to DeGóis system (web-service based)

FCT:
• October 2013 – 1st version with researcher interface
• November 2013 – 1st version with researchers bibliometric indicators
• January 2014 – 1st version of institutions interface
• Need a protocol to receive new FCT data

DGEEC:
• Need a protocol to receive new REBIDES data
Team/Thanks

Algorithms, database and interfaces:
- Sylwia Bugla (BI grant)
- Fábio Domingues (BI grant)
- João Rodrigues (BI grant)

Bibliometrics:
- Paula Pechincha (Reitoria UP)
- Elizabeth Vieira (FCT BD)
- Lara Teixeira (BI grant)

Coordinators:
- Fernando Silva (Project PI; Algorithm and implementation)
- José Ferreira Gomes (Co-PI; Bibliometrics)
Funding/Support

We are thankful to:

- FCT
- University of Porto & DCC/FCUP
- CRACS/INESC-TEC

for supporting the project.