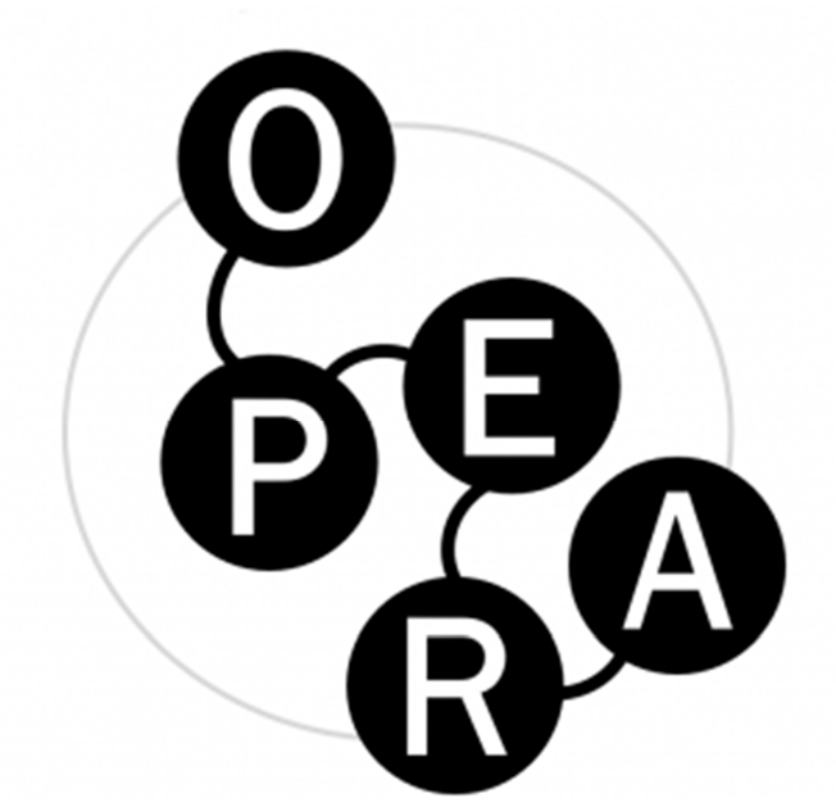
WEB OF SCIENCE & INCITES DATA FOR VIVO RESEARCH ANALYTICS PLATFORM (VIVO RAP)

Loading, enhancing, and storing



OPERA WP1 Project Report no. 1

November 29th 2019

Contributor Credits

Specified using the CRediT Taxonomy contributor roles:

Conceptualization
Ideas; formulation or evolution of overarching research goals and aims

Methodology
Development or design of methodology; creation of models

Software
Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components

Validation
Verification, whether as a part of the activity or separate, of the overall replication/ reproducibility of results/experiments and other research outputs

Formal Analysis
Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data

Investigation
Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection

Resources
Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools

Data Curation
Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse

Writing – Original Draft
Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation)

Writing – Review & Editing
Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or postpublication stages

Visualization
Preparation, creation and/or presentation of the published work, specifically visualization/ data presentation

Supervision
Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team

Project Administration
Management and coordination responsibility for the research activity planning and execution

Funding Acquisition
Acquisition of the financial support for the project leading to this publication.

Contributors and Roles

Franck Falcoz – 0000-0003-3666-5699 – fkybus@gmail.com
ROLES: Writing – Original Draft, Software, Methodology
AFFILIATION: Vox Novitas

Christina Steensboe – 0000-0002-8783-0036 – chste@dtu.dk
ROLES: Data Curation, Writing – Review and Editing
AFFILIATION: DTU – Technical University of Denmark

Nikoline Dohm Lauridsen – 0000-0002-6139-0108 – nidl@dtu.dk
ROLES: Data Curation, Writing – Review and Editing
AFFILIATION: DTU – Technical University of Denmark

Brian Lowe – 0000-0002-8143-6345 – brian@ontocale.com
ROLES: Writing – Review and Editing, Software, Methodology
AFFILIATION: Ontocale

Mogens Sandfær – 0000-0001-8436-5346 – mosa@dtu.dk
ROLES: Writing – Review and Editing, Funding Acquisition, Conceptualization
AFFILIATION: DTU – Technical University of Denmark

Karen Hytteballe Ibanez – 0000-0002-8229-0392 – kshi@dtu.dk
ROLES: Project Administration
AFFILIATION: DTU – Technical University of Denmark

TABLE OF CONTENTS

0. INTRODUCTION

The VIVO Research Analytics Platform (RAP) is an online service offering popular and standardized research analytics reports to DTU's research leadership and administration. Version 1 of the VIVO RAP was launched in 2018 with a collaboration analytics module based on co-publication analyses.

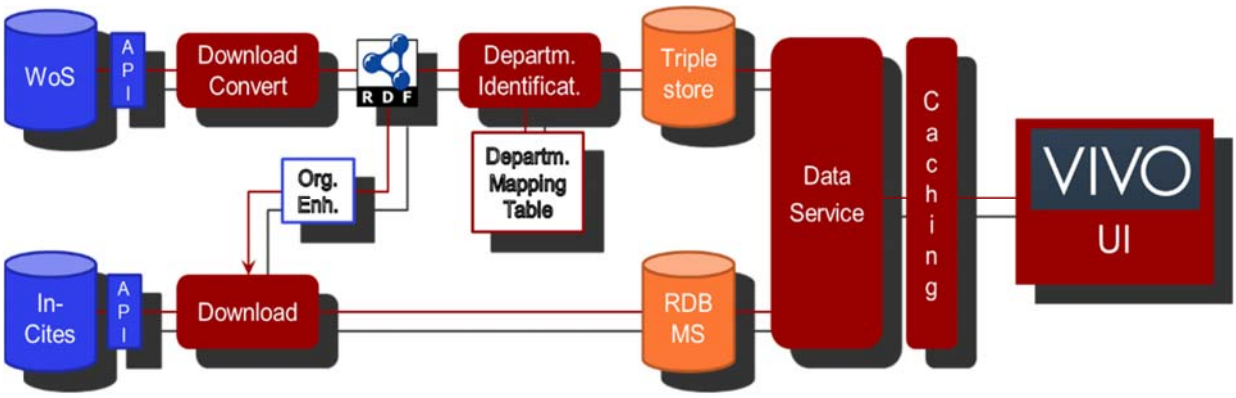
0.1 EXECUTIVE SUMMARY

This report provides an overview of the data sources used to develop the VIVO RAP as well as the data changes and enhancements completed to serve the use case of the VIVO RAP better.

1. VIVO RAP DATA SOURCES

There are two data sources for the VIVO RAP:

- **Web of Science (WoS)** for publication data pertaining to the university
- **InCites** for bibliometric indicator data pertaining to the university and all its collaboration partners



Data from WoS are stored in VIVO as RDF (Research Description Framework) triples in a triple store. The triples compose a graph of connections between discrete entities such as persons, publications, organizations, departments, and countries. These entity types and their relationships are defined in the VIVO Ontology; some minor extensions have been added to this ontology to represent details of the WoS data.

2. WOS DATA RETRIEVAL

- How is the data retrieved from WoS?

WoS data is retrieved using the SOAP API but will be updated in the coming months to use the new REST API (https://developer.clarivate.com/apis/wos) which will eventually replace the SOAP counterpart.

- Using what kind of query?

The publication data is retrieved using the unified organization search OG=(Technical University of Denmark) and by adding a timespan starting at 2007-01-01 and ending at the date of the last update of InCites.

- How often is the query executed?

Because the VIVO RAP UI depends on a synchronization between the WoS and InCites data, the WoS data is updated every time the InCites data is updated. This usually happens at the end of every month. The changes are reflected in the VIVO RAP UI a week later – after the processing, validation of the changes, and pre-caching is done.

- How to know when the InCites data is updated?

An API that returns the last update date of InCites is available:

http://api.clarivate.com/api/incites/InCitesLastUpdated/json/

Note that this API requires an API-key to work.

- Full and/or incremental updates?

The WoS data is always retrieved in full updates (approx. 40,000 records). However, after conversion, the VIVO RDF is usually added as an incremental update to the VIVO RAP. Once every three to six months, depending on changes to the data handling, all data is cleared from the VIVO RAP and a full re-load of the VIVO RDF is done.

- What is the format of the retrieved data?

The data is currently retrieved in a WoS specific XML format. When switching to the REST API, there will be a choice between XML and JSON formats. Choosing JSON would ease some of the processing, but requires extensive changes to part of the system. It has not yet been decided which of the formats will be used in the future.

- Example of the retrieved affiliation data

The following is an example of the XML for the affiliation data, taken from the WoS record: WOS:000429921000007, DOI:10.1038/s41467-018-03729-4:

```
<address_name>
    <address_spec addr_no="3">
```

```
            <full_address>
                Tech Univ Denmark, Dept Phys, DK-2800 Lyngby, Denmark
            </full_address>
            <organizations count="2">
                <organization>Tech Univ Denmark</organization>
                <organization pref="Y">Technical University of
Denmark</organization>
            </organizations>
            <suborganizations count="1">
                <suborganization>Dept Phys</suborganization>
            </suborganizations>
            <city>Lyngby</city>
            <country>Denmark</country>
            <zip location="BC">DK-2800</zip>
        </address_spec>
        <names count="1">
            <name addr_no="3" daisng_id="2404089" reprint="Y"
role="author" seq_no="2">
                <display_name>Gehring, Tobias</display_name>
                <full_name>Gehring, Tobias</full_name>
                <wos_standard>Gehring, T</wos_standard>
                <first_name>Tobias</first_name>
                <last_name>Gehring</last_name>
            </name>
        </names>
    </address_name>
```

From this data, `//organizations/organization[@pref="Y"]`(the organization enhanced name) and the `//suborganizations/suborganization` are used. The `full_address` is currently not used. Instead, the VIVO RAP relies on the interpretation that WoS makes of this address to extract the relevant organization and sub-organization.

3. WOS DATA CONVERSION

- Conversion of the WoS data to VIVO RDF

The WoS XML data is converted to RDF using python. The mapping of the WoS data to the VIVO ontology (plus a few local extensions to that ontology) was initially done with help from Clarivate Analytics.

- Example of the resulting affiliation data

After conversion to RDF, the affiliation XML shown above will result in an address and sub-organization. The address will look like this:

```
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix vivo: <http://vivoweb.org/ontology/core#> .
@prefix wos:  <http://webofscience.com/ontology/wos#> .
@prefix rap:  <http://rap.adm.dtu.dk/individual/>

rap:addr-ae65243c0a30a73ae5e1221814e1efca3
    a                       wos:Address ;
    wos:sequenceNumber      "3" ;
    rdfs:label              "Tech Univ Denmark, Dept Phys, DK-2800
    Lyngby, Denmark" ;
    wos:organizationName    "Tech Univ Denmark" ;
    vivo:relates            rap:org-technical-university-of-
denmark ;
    vivo:relates            rap:suborg-
2538f51107f3f2d55579afff7fe48cf6 ;
    vivo:relates            rap:pub-WOS000429921000007 .
```

The sub-organization generated is specific to that address:

```
rap:suborg-2538f51107f3f2d55579afff7fe48cf6
    a                       wos:SubOrganization ;
    rdfs:label              "Dept Phys, Tech Univ Denmark" ;
    wos:organizationName    "Tech Univ Denmark" ;
    wos:subOrganizationName "Dept Phys" .
```

4. WOS DATA CHANGES/ENHANCEMENTS

In general, the WoS data is converted to the VIVO RDF format without changing the information content. However, in a few cases, the information content of the data is changed/enhanced in order to serve the use cases of the VIVO RAP better:

1. Country name mapping
2. Adding missing countries
3. Choosing a single country for an organization
4. Normalizing university department names

4.1 COUNTRY NAME MAPPING

Some of the country definitions in WoS do not match the geographic data used in the VIVO ontology. Therefore, 38 WoS names are currently mapped to standard VIVO country names. Because this mapping is static, it is simply maintained in `settings.py` in a python data structure.

4.2 ADDING MISSING COUNTRIES

Some countries defined in WoS have no equivalent in VIVO's geographic data. Three countries have therefore been added to the VIVO RAP: Greenland, Macedonia, and Taiwan. These are defined in `settings.py` and added based on the RDF found in `data/countries.ttl`.

4.3 CHOOSING A SINGLE COUNTRY FOR AN ORGANIZATION

Some unified organizations have affiliations in multiple countries. Because co-publications in the VIVO RAP are counted by unified organization, this resulted in incorrect number of co-publications for some countries. The unified organizations are therefore matched to the country of their main affiliation. There are currently 103 unified organizations with such mappings. Since these are static, they are currently maintained in `settings.py` as well.

4.4 NORMALIZING UNIVERSITY DEPARTMENT NAMES

Currently only top-level organization names (unified organizations) are standardized in WoS. However, for use in the VIVO RAP where data is both filtered and grouped at the department level, this second level of affiliation was normalized.

All department name variants were extracted from WoS, simplified to lower case alphanumeric keys, and ordered by descending frequency of use in a spreadsheet. Originally, around 1,900 of these name variants were extracted, with around 60% being a long tail of variants used only once or twice.

These WoS name variants were mapped to the official DTU departments, and a number of entries, where the name was too ambiguous, were mapped to *DTU department unknown*.

WoS Department	Official DTU Department	Number of Occurrences
natl food inst	DTU Food	669
dept chem biochem engn	DTU Chemical Engineering	475
dept phys	DTU Physics	445
natl vet inst	DTU Vet	426
natl inst aquat resources	DTU Aqua	408
univ denmark	DTU department unknown	11
…		

With every monthly updates, unmapped WoS name variants are tracked, and these are manually mapped to the official DTU departments. On average, 20-30 new variants are added per update.

These department mapping spreadsheets are also used for DTU department name changes and department mergers.

4.4.1 Changes to the RDF

Given the RDF example in "3. WoS Data Conversion", the following changes are done during the department normalization phase.

All the sub-organization records related to the department of physics (DTU Physics) are removed and replaced by one common record:

```
rap:dtusuborg-dtu-physics
    a                                wos:SubOrganization ;
    rdfs:label                       "DTU Physics" ;
    wos:subOrganizationName          "DTU Physics" ;
    wos:subOrganizationNameVariant   "Dept Phys" ;
    wos:subOrganizationNameVariant   "Dept Phys Bldg" ;
    wos:subOrganizationNameVariant   "Dept Phys Bldg 307" ;
    wos:subOrganizationNameVariant   "Dept Phys Nano DTU" ;
    ... 99 more name variants ...
```

All address records pointing to these 103 different WoS sub-organizations are instead related to `rap:dtusuborg-dtu-physics`.

4.4.2 Detailed Processing

Following is a detailed description of the department standardization process with examples.

1. Load all mappings from all the past mapping spreadsheets in sequence.
2. Handle re-mapping, by checking if any of the newer mappings are actually re-mapping previous target values.

   For example, given the following mapping:

dtu cen	DTU Danchip	16	
natl ctr micro nanofabricat	DTU Danchip	11	
dtu danchip	DTU Nanolab	0	Department name change

   The first two lines will be re-mapped to:

dtu cen	DTU Nanolab	16	Overwritten from 'DTU Danchip'
natl ctr micro nanofabricat	DTU Nanolab	11	Overwritten from 'DTU Danchip'

3. Find all addresses related to DTU.
4. Find all sub-organizations related to these addresses.
5. Extract sub-organization names from these sub-organization records and map them to lower case alpha-numeric keys e.g.

Dept Biotechnol & Biomed Tech	dept biotechnol biomed tech
Dept Energy Conversion & Storage	dept energy conversion storage
Atmospher Environm Sect	atmospher environm sect

6. Check these keys against the current mapping. If a mapping already exists, it is used as the new sub-organization name; otherwise, the key is added to the list of new sub-organizations to be mapped manually.

5. INCITES DATA RETRIEVAL AND STORAGE

In addition to the data retrieved from WoS for the Technical University of Denmark, bibliometric indicators for entire organizations are retrieved from InCites for the Technical University of Denmark and all the collaborating organizations.

- How is the data retrieved from InCites?

A number of indicators are retrieved from InCites using a query of the unified organization names e.g. OG=(Technical University of Denmark)

- How often is the query executed?

Since the WoS and InCites data should stay synchronized and InCites has the longest update period (monthly), the update for both WoS and InCites runs as soon as the API described in "2. WoS Data Retrieval" returns a new update date.

- Full and/or incremental updates?

Full updates are always done, as there is no option for getting partial updates. The bibliometric indicators are also updated as a full update to MySQL. These indicators are stored in MySQL and used directly during the creation of the VIVO RAP collaboration reports.