

# Data integration for paleoenvironmental and archaeological GIS based analysis

Christian Willmes<sup>1</sup>, Daniel Becker, Georg Bareth

GIS & RS Group, Institute of Geography, University of Cologne

CAA 2015 - Keep the Revolution going  
Siena, Italy - March 31st, 2015

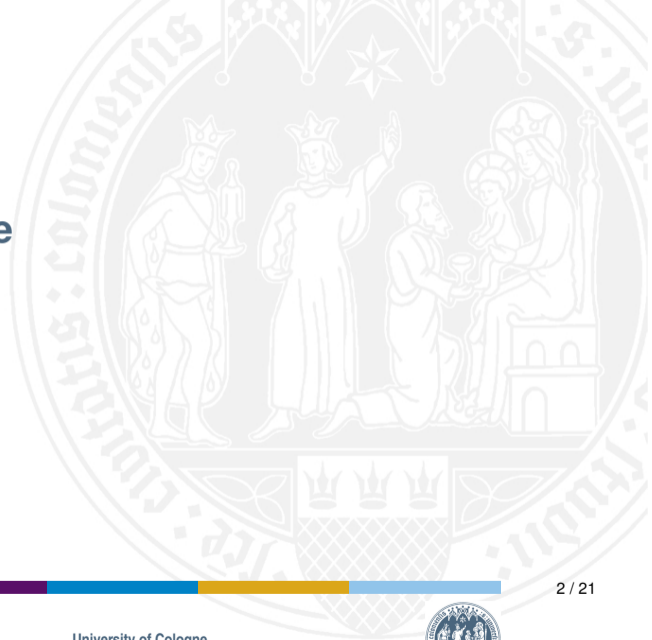
---

<sup>1</sup>c.willmes@uni-koeln.de



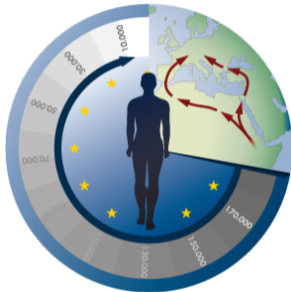
# Outline

1. Introduction
2. Tools and infrastructure
3. Data integration
4. Data discovery
5. Data transformation
6. Data analysis (GIS)



# Introduction

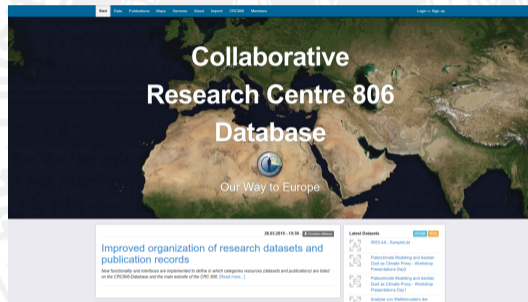
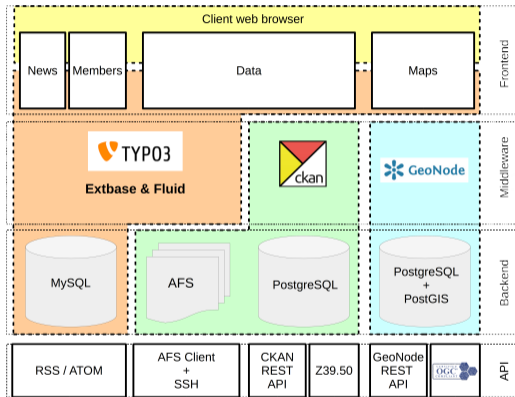
Collaborative Research Centre 806



- Interdisciplinary research project
- Concerns the history of mankind, and in particular the migration/expansion process from africa to europe
- Disciplines of archaeology, geosciences and cultural sciences
- Planned for an overall duration of 12 years (3 x 4 year terms), since 7/2013 in second phase
- More information about the project in detail is available at: <http://www.sfb806.de/>

# Introduction

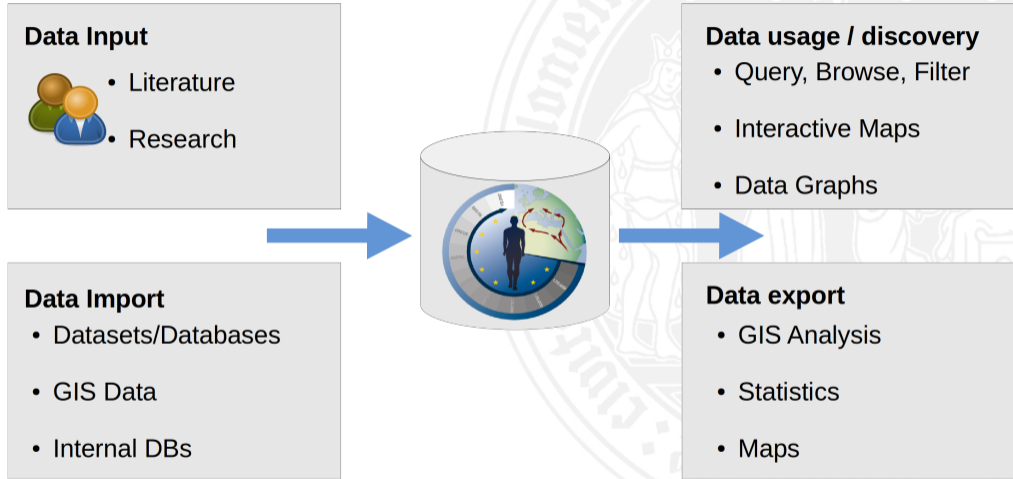
## CRC806-Database



<http://crc806db.uni-koeln.de/>



# Collaborative Research Database



# Semantic Mediawiki

- Allows to add **structured data** to a Wiki
- Allows **collaborative editing** of structured Data
- Allows **complex queries**
- Allows **export and display** of query results in many formats



# Data model development

## Prototyping approach

Implementation  
& Updating



Requirements  
& Feedback



Data model & Prototype



# Data model development

## Mobo Framework

- <https://github.com/Fannon/mobo>
- Allows to develop the data model in **JSON-Schema** notation
- Model adjustments are applied centrally (normally many things to consider in SMW)
- Datamodel can be managed in **VCS like GIT**
- Same model can easily be **deployed to many SMW** instances

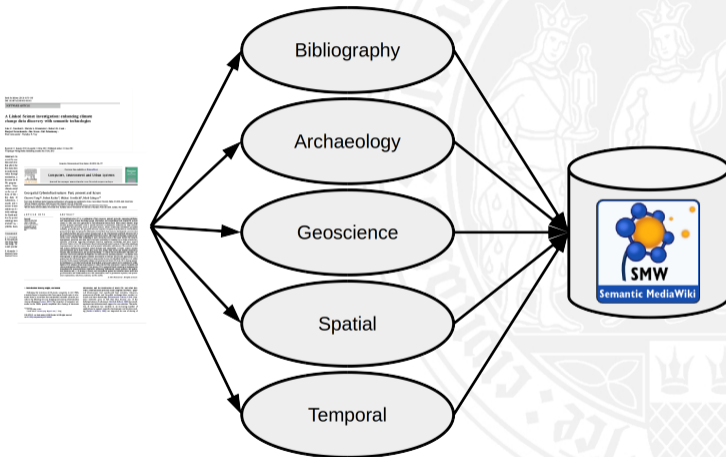
The logo for the Mobo framework, consisting of the word "mobo" in a white, lowercase, sans-serif font centered within a solid black rectangular box.





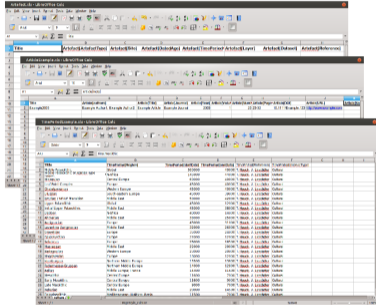
# Data integration

## Data from literature workflow

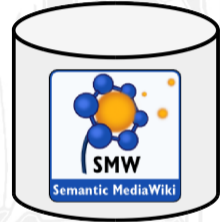


# Data integration

## Spreadsheet Template Import



Data Transfer  
(Mediawiki Ext.)

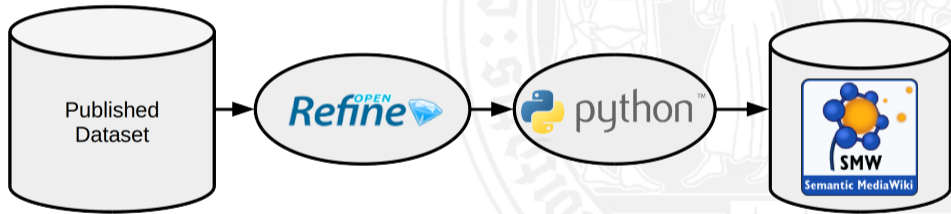


Spreadsheets with fixed schemas (templates) of SMW classes.



# Data integration

## Integration of datasets workflow



# Data cleaning using OpenRefine

Google refine INQUA\_View\_all\_fields.xls Permalink

Open... Export Help

Facet / Filter Undo / Redo 2

21499 rows Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next > last »

Refresh Reset All Remove All

x 9.0 change Cluster

2289 choices Sort by: name count

Early Upper Palaeolithic, Gravettoid 5

Early Upper Palaeolithic, Jermanovician 2

Early Upper Palaeolithic, Late Middle Palaeolithic 1

Early Upper Palaeolithic, Lincombian-Ranisian-Jerzmanowician technocomplex 1

Early Upper Palaeolithic, lower part of the Lower Aurignacian, Uluzzian 10

Early Upper Palaeolithic, Mousterian 3

x 9.0 change reset



	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0
1.	verified	id	g_sitename	g_layer_id	g_town	g_province	g_country	g_coord_long	g_coord_lat	cu_stage	cu_anthr_remains	s_presumed_mi
2.	1	3	La Vina		Manzaneda	Asturias	Spain	5.83	43.314	Final Magdalenian		
3.	1	4	La Vina	V & VI	Manzaneda	Asturias	Spain	5.83	43.314	Solutrean		
4.	1	6	La Vina	III	Manzaneda	Asturias	Spain	5.83	43.314	Upper Magdalenian		
5.	1	7	Oscura de Perin			Asturias	Spain	-5.96306	43.40306	Chatelperronian		
6.	1	8	Oscura de Perin			Asturias	Spain	-5.96306	43.40306	Solutrean		
7.	1	9	Oscura de Perin			Asturias	Spain	-5.96306	43.40306	Azilian		
8.	1	10	Pena Ferran			Asturias	Spain	-5.3630555555555555	43.313888888888888	Upper Magdalenian, Solutrean/Magdalenian		
9.	1	11	Salumula		Aviao	Asturias	Spain	-6.35806	43.29644	Solutrean		
10.	1	12	Sulamula			Asturias	Spain	-5.1419444444444444	43.47	Solutrean		
11.	1	13	Los Canes		Arangas, Cabrales	Asturias	Spain	-4.85	43.3167	Late Magdalenian		
12.	1	14	Los Azules	3a-3f	Cangas de Onis	Asturias	Spain	-5.1324999999999999	43.0667	Azilian		
13.	1	15	Cova Rosa		Riba de Sello	Asturias	Spain	-5.11	43.44	Solutrean		
14.	1	16	Cova Rosa		Riba de Sello	Asturias	Spain	-5.11	43.44	Upper Magdalenian		

<http://openrefine.org/>



# Data import via Python Script

```
~/ownCloud/Research/Vortrag/2015_CAA/Img/INQUAtoSMW.py - Sublime Text (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help
INQUAtoSMW.py
1  #!/usr/bin/python
2
3  import xlrd
4  import xml.etree.cElementTree as ET
5  import itertools, collections
6
7  def consume(iterator, n):
8      collections.deque(itertools.islice(iterator, n))
9
10 #Excel Definitions
11 excelFile = xlrd.open_workbook("INQUA_sanitized.xls")
12
13 #XML init
14 root = ET.Element("Pages")
15
16 iterator = range(10001, sheet.nrows-1).__iter__()
17 for i in iterator:
18     row = sheet.row_values(i)
19
20     sname = unicode(sanitize_name(row[2]))
21
22     Site = ET.SubElement(root, "Page")
23     Site.set("Title", sname)
24
25     [...]
26
27 #write output
28 tree = ET.ElementTree(root)
29 tree.write("INQUA.xml")
30
```



# Data discovery

## Example: Query Sites by TimePeriod and Region

```
[[[-Site::<q> [[Category:Artefact]] [[TimePeriod::Solutrean]]</q>]] [[Country::France]]
```

**Semantic search**

Query	Additional data to display (add one property name per line)
<pre>[[[-Site::&lt;q&gt; [[Category:Artefact]] [[TimePeriod::Solutrean]]&lt;/q&gt;]] [[Country::France]]</pre>	?Coordinates

Format as:  For a detailed description, please visit the [KML help page](#).

Sorting  
[\[Add sorting condition\]](#)

limit: <input type="text" value="200"/> The maximum number of results to return	offset: <input type="text" value="0"/> The offset of the first result	link: <input type="text" value="all"/> Show values as links
sort: <input type="text"/> Property to sort the query by	order: <input type="checkbox"/> descending <input type="checkbox"/> desc <input type="checkbox"/> asc <input type="checkbox"/> ascending <input type="checkbox"/> rand <input type="checkbox"/> random Order of the query sort	headers: <input type="text" value="show"/> Display the headers/property names
mainlabel: <input type="text"/> The label to give to the main page name	intro: <input type="text"/> The text to display before the query results, if there are any	outro: <input type="text"/> The text to display after the query results, if there are any
searchlabel: <input type="text"/> Text for continuing the search	default: <input type="text"/> The text to display if there are no query results	text: <input type="text"/> The text associated with each page. Overridden by the additional queried properties if any.
title: <input type="text"/> The default title for results	linkabsolute: <input checked="" type="checkbox"/> Should links be absolute (as opposed to relative)	pageinltext: <input type="text" value="View page \$1"/> The text to use for the links to the page, in which \$1 will be replaced by the page title

[Find results](#) | [Hide query](#) | [Show embed code](#) | [Querying help](#)



# Data discovery

## Example: Query Sites by TimePeriod and Region

Form fields for map configuration:

- Polgons to display:
- Circies to display:
- Rectangles to display:
- wmsoverlay:  Use a WMS overlay
- maxzoom:  The maximum zoom level
- minzoom:  The minimum zoom level
- copycoords:  Show a dialog when clicking on a location from which its coordinates may be copied
- static:  Make the map static

Find results | Hide query | Show embed code | Querying help

Map showing query results (red pins) across Europe, primarily in France and Italy. The map includes navigation controls and a scale bar. Text on the map includes: Results 1 - 22, Next, (20 | 50 | 100 | 250 | 500), Previous, Results 1 - 22, Next, (20 | 50 | 100 | 250 | 500), Map data ©2015 Basarsoft, GeoBasis DE/BKG (©2009), Google, basado en BCN/IGN España, 100 km, Terms of Use.

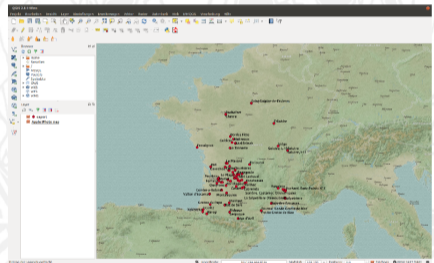
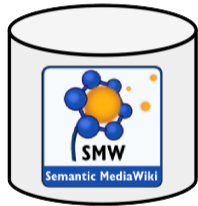
Footer: Privacy policy | About SMW | Disclaimers | Powered by InetSoft | Connected by Semantic MediaWiki





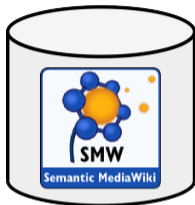
# Export data to GIS

SMW Queryresult in KML result format



# GIS Analysis

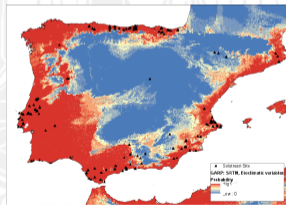
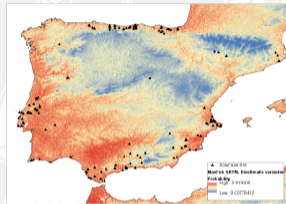
## Niche- or Species distribution- Modelling



Refine<sup>OPEN</sup>

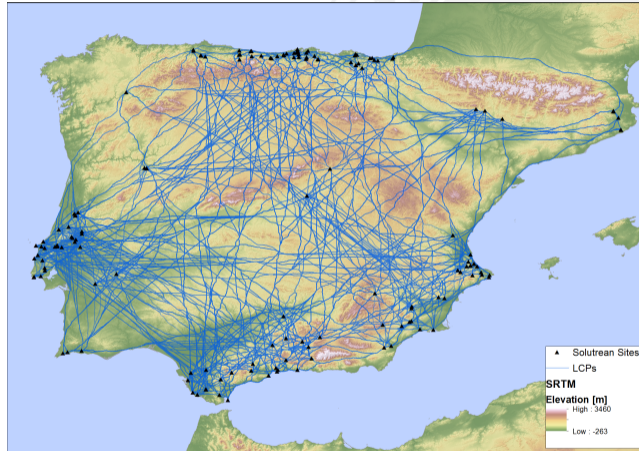


open  
modeller



# Data analysis

## Networkanalysis based on LCPs



# Conclusions and Outlook

- **Integrated database of diverse information**
- **Development of an "tailor-made" data model**
- **Mapping the data model to existing Vocabularies**
- **Publish the database as Linked Data**



# Thank you very much!

eMail: [c.willmes@uni-koeln.de](mailto:c.willmes@uni-koeln.de)

twitter: [@cwillmes](https://twitter.com/cwillmes)

<http://crc806db.uni-koeln.de>

This research was conducted within the Collaborative Research Centre 806 ([www.sfb806.de](http://www.sfb806.de)) funded by the German Research Foundation ([www.dfg.de](http://www.dfg.de)).

