INDICATE

INDICATE
International Network for a Digital Cultural Heritage e-Infrastructure

Introduction to geocoded cultural content (GCC): framework, use cases, geoparsing

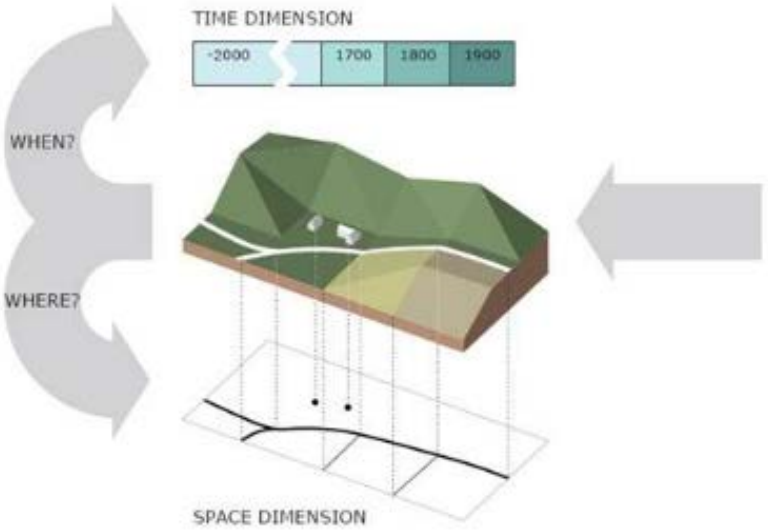
Franc J. Zakrajsek
scientific coordinator of Indicate for Slovenia, Slovenia

International workshop: Geocoded cultural content
Ljubljana, 7th February 2012

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

What is GCC ?

cultural objects and events



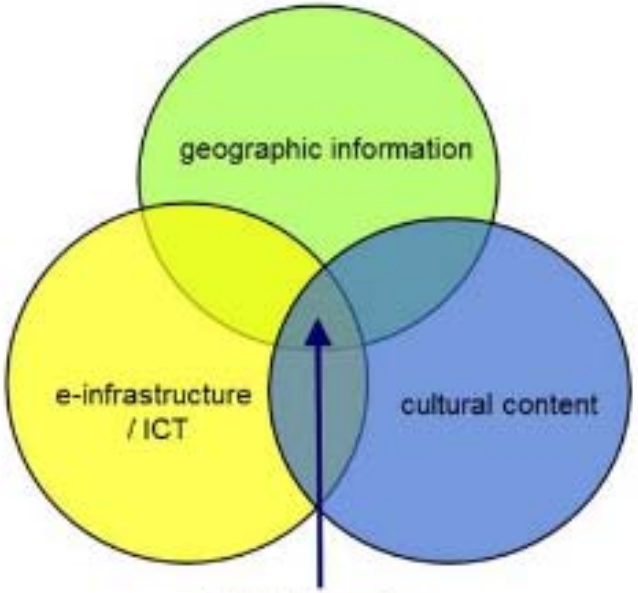
books
exibitions
concerts
archeological sites
museum objects
paintings
monuments
maps
theater shows
festivals
architectural buildings

Geographic location is one of the most important aspect of information for every cultural heritage item. A formalized location attribute (e.g. geocode or geographical coordinates) will significantly enhance the power of searching, visualization, analysis of the content.

INDICAT
International Network for Digital Cultural Heritage e-Infrastructure

What we are doing ?

The case study Geocoded cultural content: first part reviews the current approaches and new R&D on geocoding of cultural content in digital libraries, cultural tourism, heritage, e-learning, living arts and other cultural areas. Main area of the research will identify the possibilities and benefits of using e- infrastructure. The focus will be primarily on cloud and grid computing and data infrastructures when dealing with geocoded digital cultural content. The last part of the research provides and summarizes the testing of geoparsing and geotagging e-services in digital culture and recommendations for content providers.



digital libraries
cultural tourism
heritage research
e-learning
living arts
social networks

INDICAT
International Network for Digital Cultural Heritage e-Infrastructure

Concepts and framework of GCC

contex of use

spatial accuracy

linking open data

digital object types

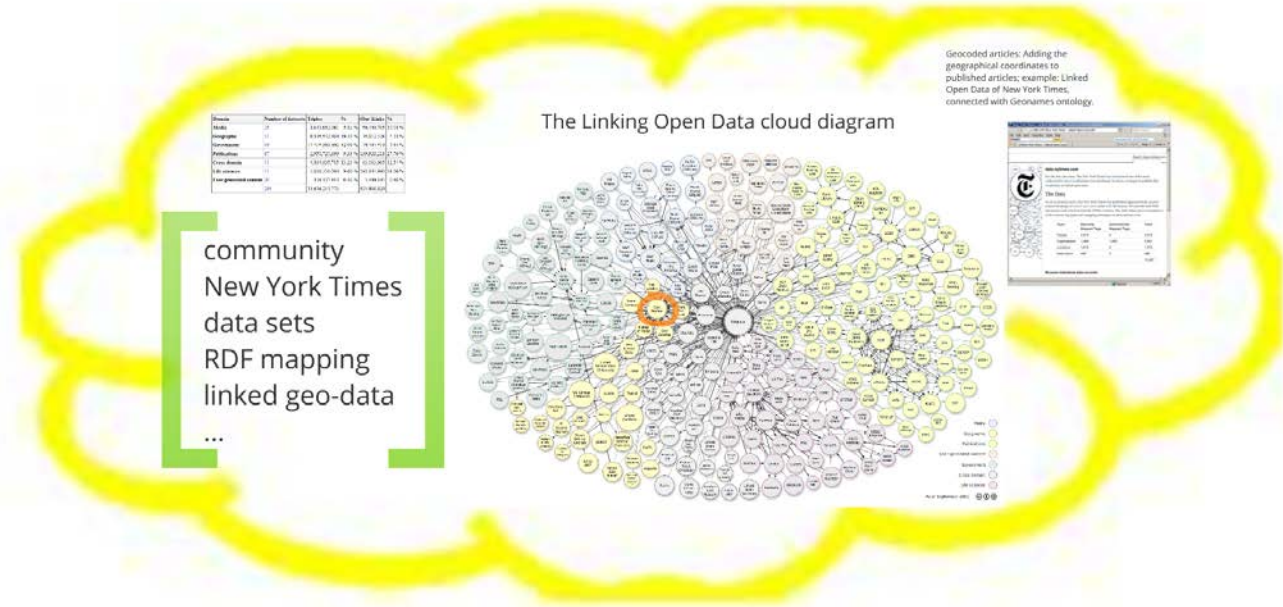
standards

geo feature types

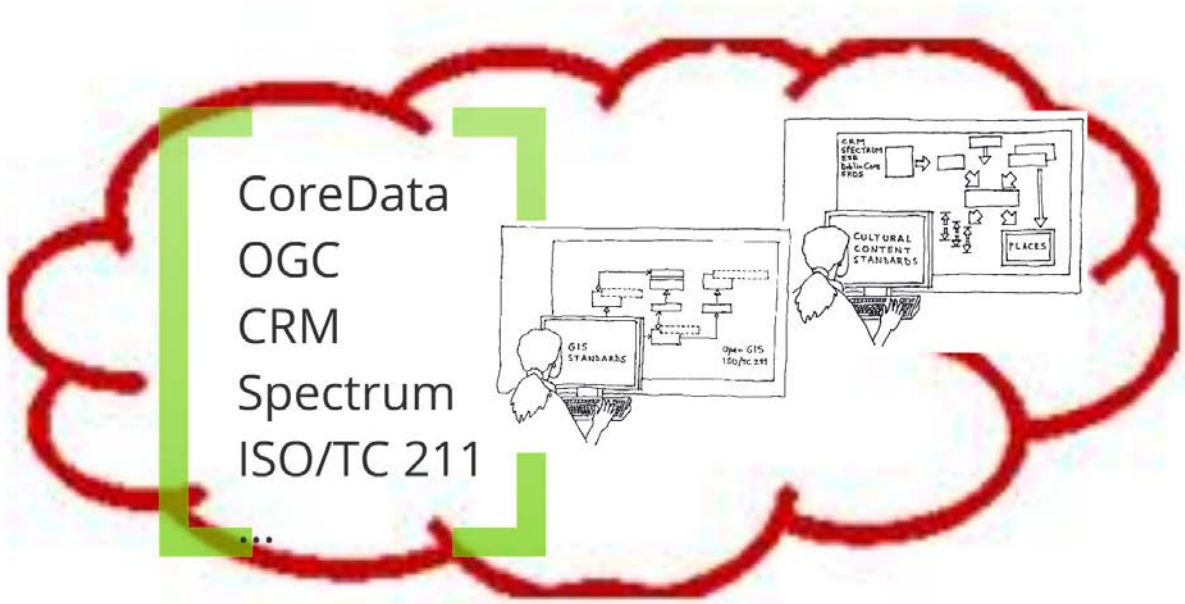
geocoding

devices

e-infrastructure



linking open data


CoreData
OGC
CRM
Spectrum
ISO/TC 211
...
GIS STANDARDS
CULTURAL CONTENT STANDARDS
PLACES

standards

INDICAT
International Network for Digital Cultural Heritage e-Infrastructure

Archeological and architectural GCC

Architectural / archeological heritage

Definition
Architectural and archaeological heritage refers to a place, locality, natural landscape, settlement area, architectural complex, archaeological site, or standing structure from inventories, management, restoration, ...

Examples
National Heritage List for England
National Heritage Register Netherlands
National Register of Denmark
Heritage Register Bayern - Nürnberg

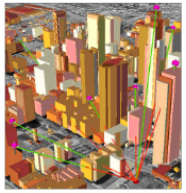
Geographical information
Geographic information in archaeological / architectural sectors is used when capturing data, management the repositories, and processing and displaying data on the maps. The level of detail goes to individual site or object.

e-Infrastructure
Appropriate tasks for grid computing:

- Risk scenarious simulations
- 3D visualisation
- Spatial statistics
- Spatial analyses

Potentials of use grid computing for caching

Steps	Software and hardware	Estimated time
Experiment Cashing area: 152 km2 scale 1:76 (approx. 2D 1:1000) Tiles: 512x512 pixels (finally 104.000 tiles (3/4 tilov))	ArcGIS Server 2x E5450 3GHz (8 threads) 32GB memory	Cashing time: 77 minut.
Generalization for the world World mainland: 148.429.000 km2, million times larger area than in experiment	ArcGIS Server 2x E5450 3GHz (8 threads) 32GB memory	Estimated 77 milions minutes or 146 years
Google	• In 2002; upwards of 15,000 servers • A 2005 estimate by Paul Strassmann has 200,000 servers claimed this number to be upwards of 450,000 in 2006 900.000 (2011)	



INDICAT
International Network for Digital Cultural Heritage e-Infrastructure

Registers of immovable CH

National Heritage List for England

Online database enables searching for listed buildings, scheduled monuments, protected wreck sites, registered parks and gardens, registered battlefields in classic way and on a map.

National Heritage Register Netherlands

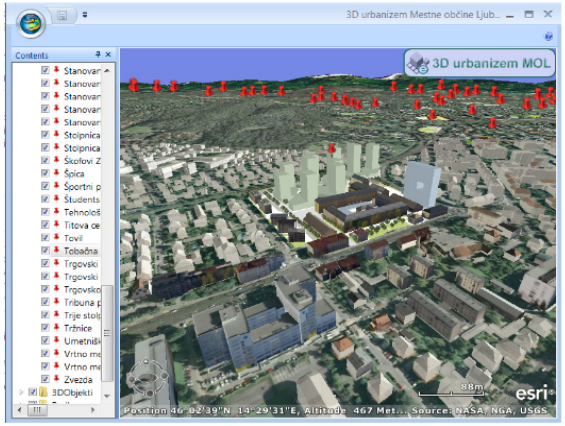
Online heritage register of national monuments (over 60.000) is split per province is searchable in classic way and on a map.

National Register of Sites and Monuments Denmark

Online register of all known sites, monuments and archaeological finds (165,000 sites including shipwrecks and submarine Stone Age settlements)

German Heritage Register Bayern - Nürnberg

Nürnberg displays part of the Bavarian monument list. Monument searching and viewing on a map is part of a city plan internet application.



INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Registers of immovable CH

National Heritage List for England

Online database enables searching for listed buildings, scheduled monuments, protected wreck sites, registered parks and gardens, registered battlefields in classic way and on a map.

National Heritage Register Netherlands

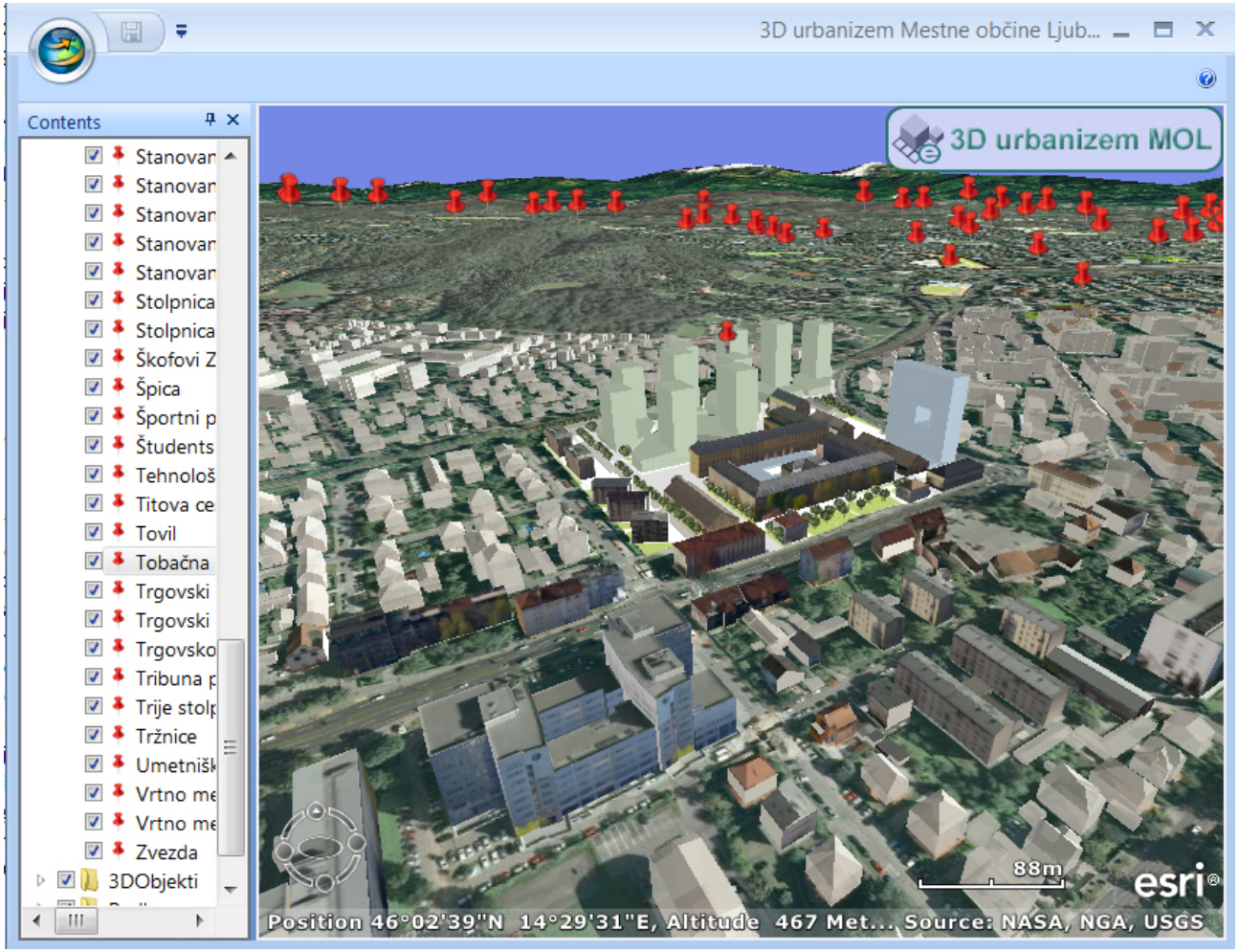
Online heritage register of national monuments (over 60.000) is split per province is searchable in classic way and on a map.

National Register of Sites and Monuments Denmark

Online register of all known sites, monuments and archaeological finds (165,000 sites including shipwrecks and submarine Stone Age settlements)

German Heritage Register Bayern - Nürnberg

Nürnberg displays part of the Bavarian monument list. Monument searching and viewing on a map is part of a city plan internet application.


3D urbanizem Mestne občine Ljub...
Contents
Stanovan
Stanovan
Stanovan
Stanovan
Stanovan
Stolpnica
Stolpnica
Škofovi Z
Špica
Športni p
Študents
Tehnološ
Titova ce
Tovil
Tobačna
Trgovski
Trgovski
Trgovsko
Tribuna p
Trije stolp
Tržnice
Umetniš
Vrtno me
Vrtno me
Zvezda
3DObjekti
3D urbanizem MOL
88m
esri
Position 46°02'39"N 14°29'31"E, Altitude 467 Met..., Source: NASA, NGA, USGS

Potentials of use grid computing for caching

Steps	Software and hardware	Estimated time
Experiment Cashing area: 152 km2 scale 1:76 (approx. 2D 1:1000) Tiles: 512x512 pixels (finally 104.000 tiles (3/4 tilov))	ArcGIS Server 2x E5450 3GHz (8 threads) 32GB memory	Cashing time: 77 minut.
Generalization for the world World mainland: 148.429.000 km2, million times larger area than in experiment	ArcGIS Server 2x E5450 3GHz (8 threads) 32GB memory	Estimated 77 milions minutes or 146 years
Google	• In 2002; upwards of 15,000 servers • A 2005 estimate by Paul Strassmann has 200,000 servers claimed this number to be upwards of 450,000 in 2006 900.000 (2011)	

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Digital libraries GCC

Use cases of digital libraries

Definition
Digital library is a collection of digital content from libraries, archives, museums and other cultural institutions. It contains internal collections in e.g. museum or in certain branch e.g. movable heritage and resides at national level, european level and world level.

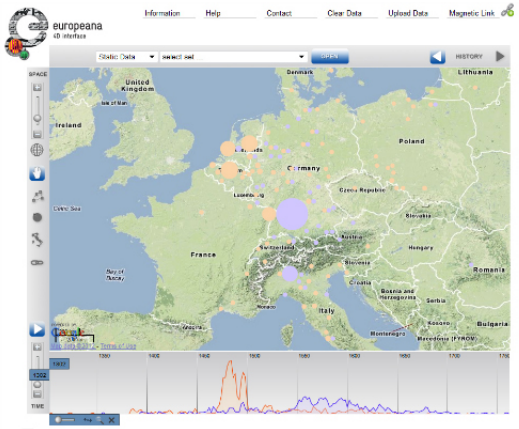
Geographical information
Support for visualization, processing geographic information about digital cultural objects. It includes also geocoded historical maps.

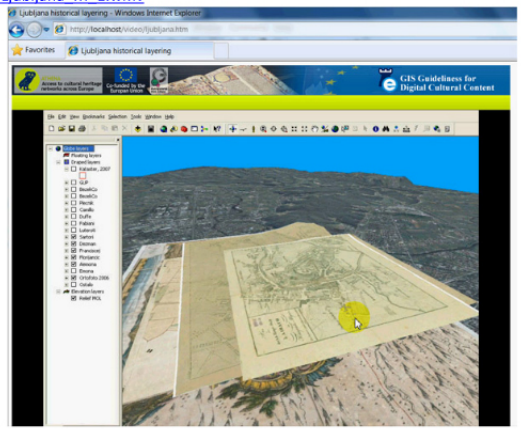
Examples

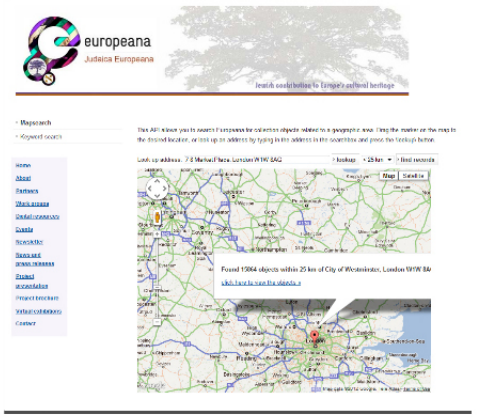
Europeana globe:	http://www.europeanaglobe.eu/
Europeana:	http://www.europeana.eu/portal/
arXiv:	http://www.arXiv.org
American Memory:	http://memory.loc.gov/ammem/index.html
dLib:	http://www.dlib.si/
Europeana 4D:	http://wp1187670.wp212.webpack.hosteurope.de/e4d/
Juidaica:	http://www.judaica-europeana.eu/map
3d historical maps:	http://indicate.situla.org/indicate/Ljubljana_M_1.wmv

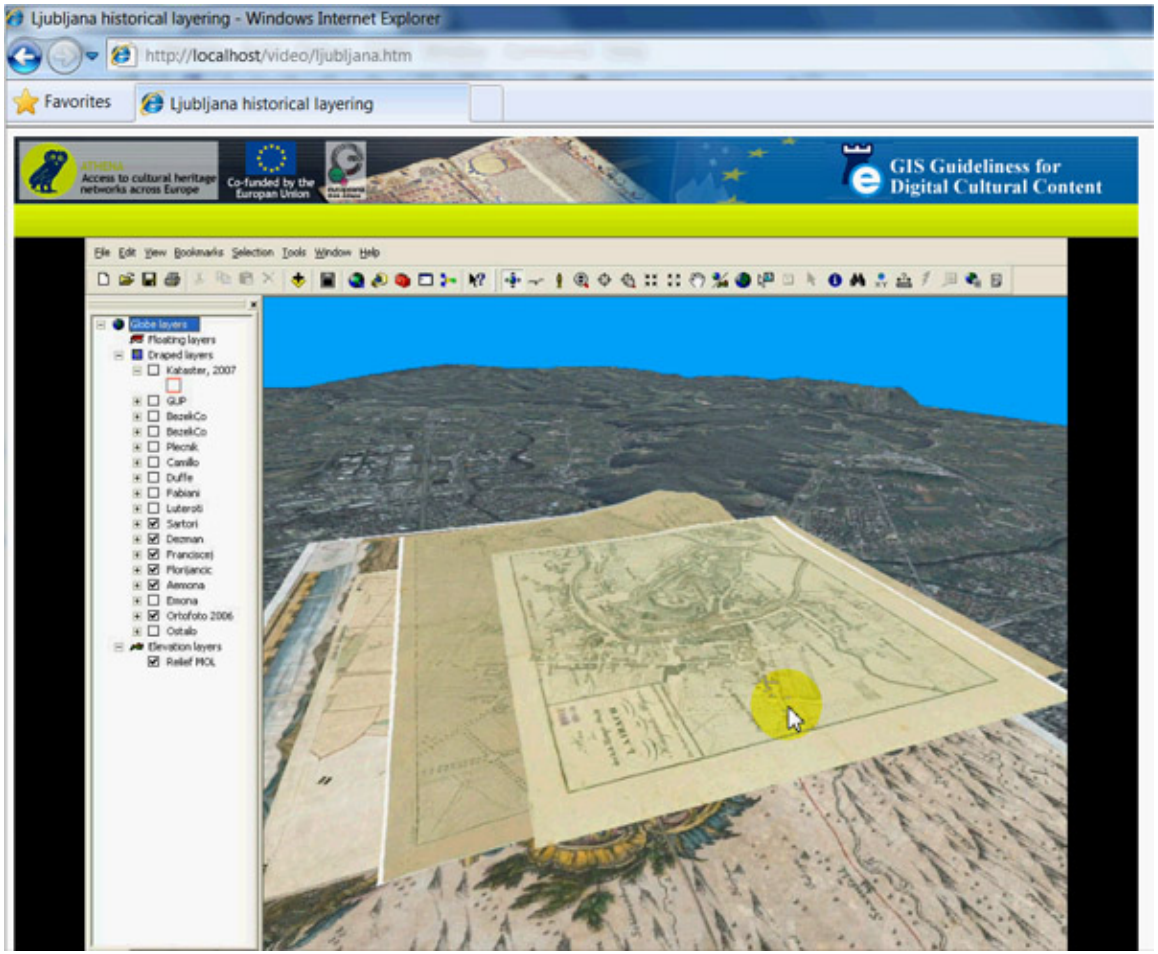
e-Infrastructure
Appropriate for grid computing:

- Digital object of content providers
- GIS, ICT
- Bussines models (“from the street”, established e.g. ECC, NREN, ...)








Ljubljana historical layering - Windows Internet Explorer
http://localhost/video/ljubljana.htm
Favorites
Ljubljana historical layering
ATHENA
Access to cultural heritage networks across Europe
Co-funded by the European Union
GIS Guideliness for Digital Cultural Content
File Edit View Bookmarks Selection Tools Window Help
Globe layers
Floating layers
Draped layers
Kataster, 2007
GUP
BezekCo
BezekCo
Plecnik
Camillo
Duffe
Fabiani
Luteroti
Sartori
Dezman
Franciscej
Florijancic
Aemona
Emona
Ortofoto 2006
Ostalo
Elevation layers
Relief PKL

e-Infrastructure

Appropriate for grid computing:

- Digital object of content providers
- GIS, ICT
- Bussines models (“from the street”, established e.g. ECC, NREN, ...)

INDICATE

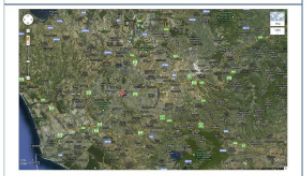
International Network for Digital Cultural Heritage e-Infrastructure

Geoparsing

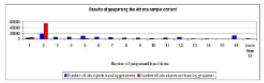
Geoparsing ?

Geoparsing is the process of assigning geographic coordinates to textual words and phrases (e.g. "The author has been born in Rome"). Geoparsing is capable of identifing ambiguous references in unstructured content. Geoparsed features can then be mapped and entered into a geographic information system. A **geoparser** is a piece of software or a (web) service that helps in this process.

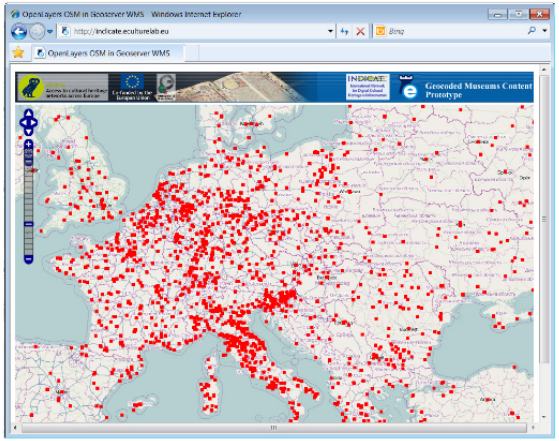
How it works ? /3



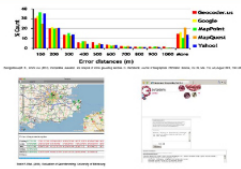
Testing with Athena data /2



Results are for the sample of the Athena content (3,87%) The first input for geoparser is whole Lido object and the second input all «place» tags included in the Lido object. The geoparser found at least one coordinates in 60,37% of Lido objects and did not fount any coordinates in 39,63% of Lido objects. The exactness of the coordinates found in Geonames gazetter is not the subject of the analysis.



State of the art



Purpose of testing

- Could we find out geographical coordinates from the textual metacata of the certain digital content ?
- What strategies and geoparsing services could we use for geoparsing ?
- What percentage of the content could be geocoded in this way, at the best?
- For what purpose / services could we use the geoparsed geographical coordinates (spatial accuracy) ?
- To plan the real production of geoparsing

How it works ? /1

Methods of testing

- The Europeana Geoparser v 1.0 Beta is used
- Input for testing: 4.082.919 LIDO objects in XML format
- Analysis the geoparsing results
- Verification the results on the map

Spatial accuracy

How it works ? /2

Testing with Athena data /1

The majority of Lido objects has at least one «Place» tag: 75,53% and 14,83% of the objects do not have «Place» tag. The analysis did not include the analysis of the «Place» tag itself. If the places exist or their syntax was not the subject of the analysis.

Conclusions

- Use geoparsing for upper level of LOD (Level Of Detail)
- If there are small town or villages they were seldom found, inclusion of national register of geographic names is strongly suggested
- Use geoparsing for validation when existings coordinates are correct
- For current locations use exact coordinates of museum or other cultural memory institution instead of geoparsing
- Use geotagging instead of geoparsing where possible
- Assigning the geographic coordinates as part of documentation process where possible

e-infrastructure

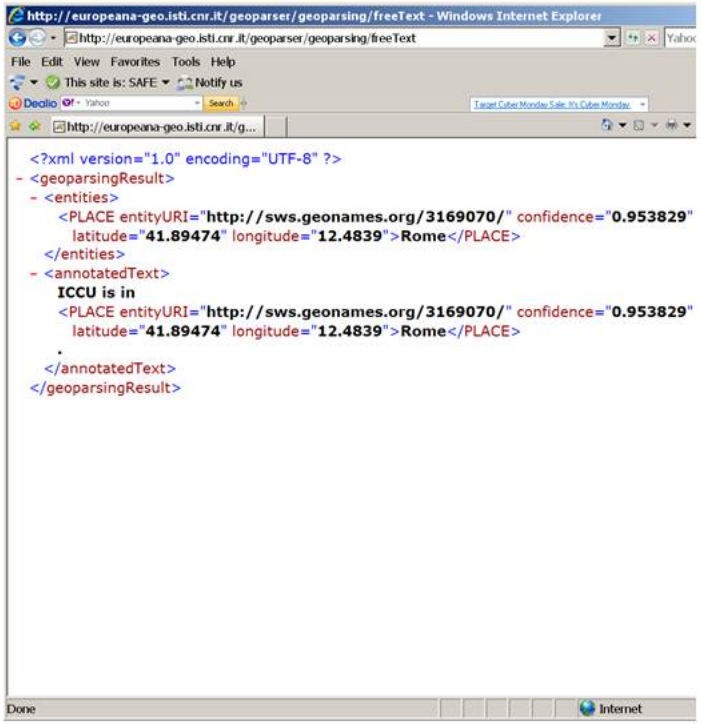
- Appropriate for grid computing
- Natural language processing (NLP)
- Use of local Gazetteers and other data sources

Geoparsing ?

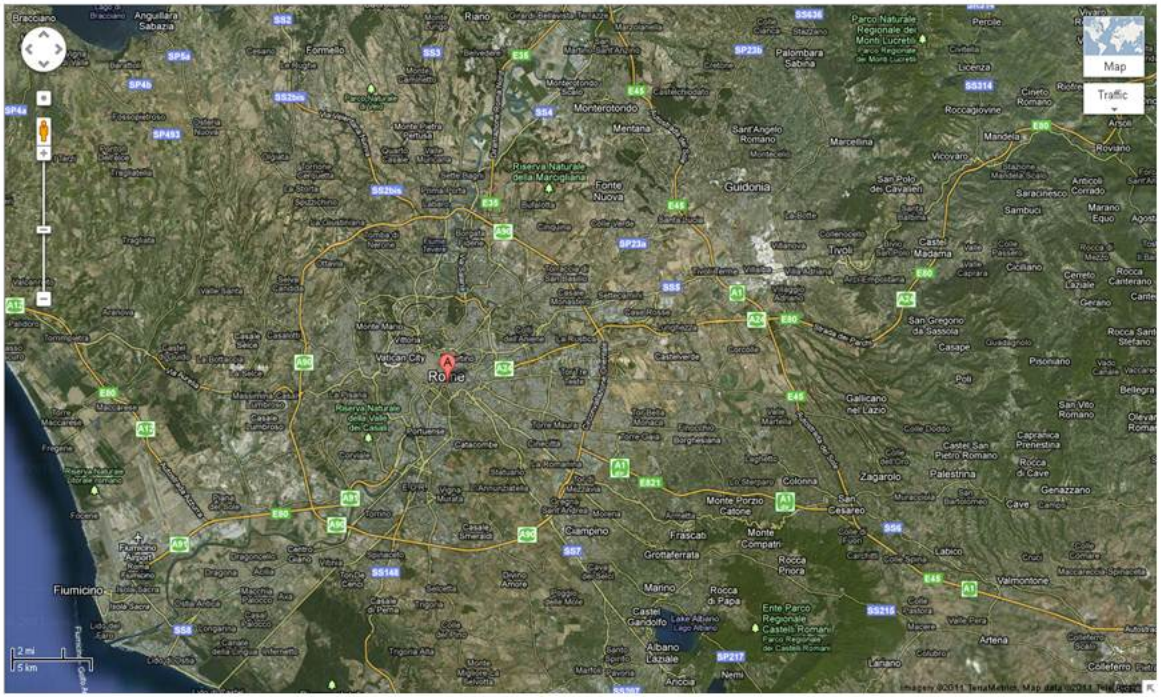
Geoparsing is the process of assigning geographic coordinates to textual words and phrases (e.g. “The author has been born in Rome”). Geoparsing is capable of handling ambiguous references in unstructured content. Geoparsed features can then be mapped and entered into a geographic information system. A **geoparser** is a piece of software or a (web) service that helps in this process.

How it works ? /1

How it works ? /2



How it works ? /3



Purpose of testing

- Could we find out geographical coordinates from the textual metadata of the certain digital content ?
- What strategies and geoparsing services could we use for geoparsing ?
- What percentage of the content could be geocoded in this way, at the best?
- For what purpose / services could we use the geoparsed geographical coordinates (spatial accuracy) ?
- To plan the real production of geoparsing

Methods of testing

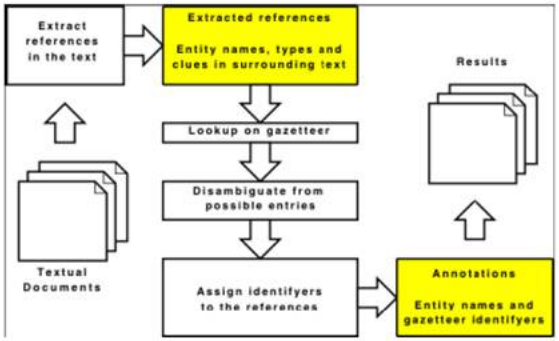
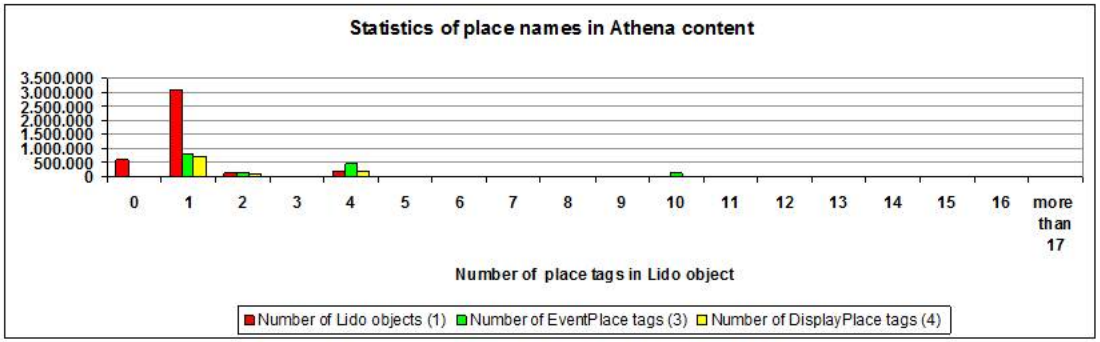


Figure 1. Typical approach for geo-parsing text

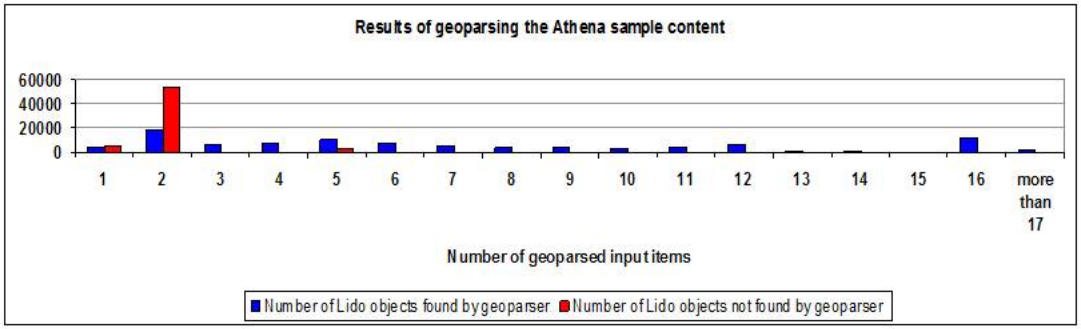
- The Europeana Geoparser v 1.0 Beta is used
- Input for testing: 4.082.619 LIDO objects in XML format
- Analysis the geoparsing results
- Verification the results on the map

Testing with Athena data /1



The majority of Lido objects has at least one »Place« tag: 75,53% and 14,83% of the objects do not have »Place« tag. The analysis did not include the analysis of the »Place« tag itself. If the places exist or their syntax was not the subject of the analysis.

Testing with Athena data /2

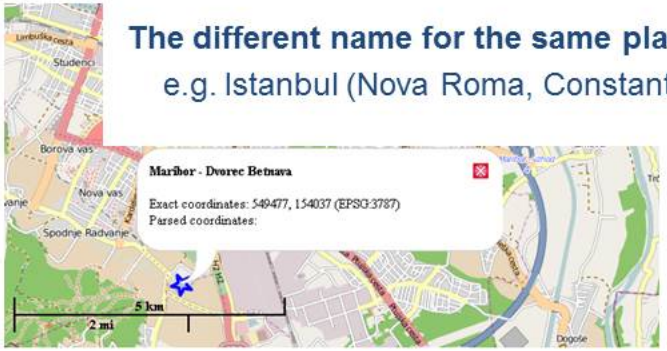


Results are for the sample of the Athena content (3,87%) The first input for geoparser is whole Lido object and the second input all »place« tags included in the Lido object. The geoparser found at least one coordinates in 60,37% of Lido objects and did not fount any coordinates in 39,63% of Lido objects. The exactness of the coordinates found in Geonames gazetter is not the subject of the analysis.

Spatial accuracy

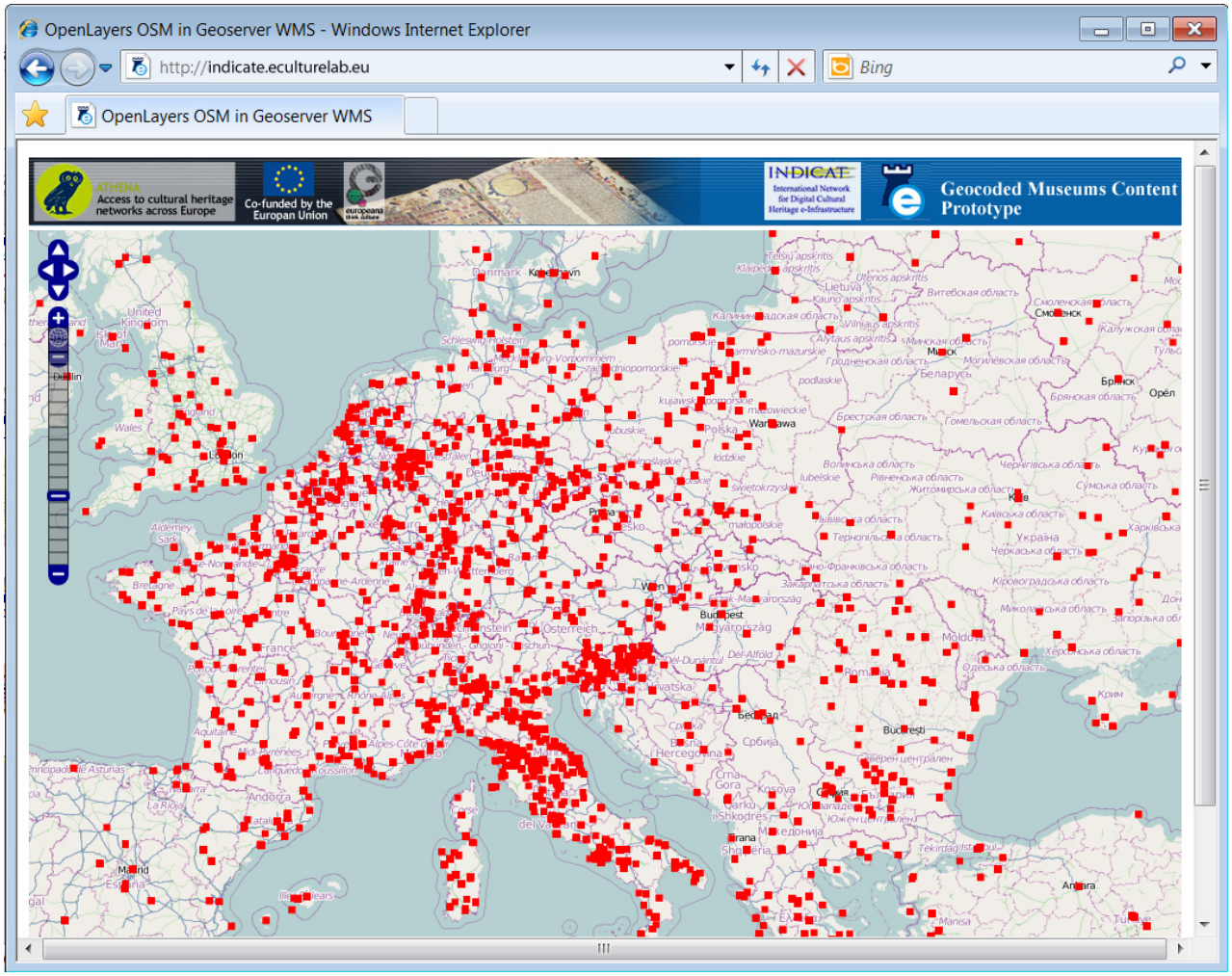
The same geographic name for different places

e.g. “Paris” adresses 93 places



The different name for the same place

e.g. Istanbul (Nova Roma, Constantinople, Tsargrad, ...)


OpenLayers OSM in Geoserver WMS - Windows Internet Explorer
http://indicate.eculturelab.eu
Bing
OpenLayers OSM in Geoserver WMS
ATHENA
Access to cultural heritage networks across Europe
Co-funded by the Europan Union
europeana
INDICATE
International Network for Digital Cultural Heritage e-Infrastructure
Geocoded Museums Content Prototype

e-infrastructure

- Appropriate for grid computing
- Natural language processing (NLP)
- Use of local Gazetteers and other data sources

Conclusions

- Use geoparsing for upper level of LOD (Level Of Detail)
- If there are small town or villages they were seldom found, inclusion of national register of geographic names is strongly suggested
- Use geoparsing for validation when existings coordinates are correct
- For current locations use exact coordinates of museum or other cultural memory institution instead of geoparsing
- Use geotagging instead of geoparsing where possible
- Assigning the geographic coordinates as part of documentation process where possible

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Knowledge Café

1. table: Archaeological / architectural heritage and GIS
Facilitator: Matteo Lorenzini
Expected attendee: archaeological and architectural institutions (10+)
Questions:
- identification of additional use cases (regularly operating, research, planed) (additional form) navigation…
- benefits and weakness of open source
- effective browsing of 3D cities
- geographical coordinate systems
- the needs for grid and cloud computing (restoration, 3D rendering, cashing, …)

2. table: Libraries and GIS
Facilitator: Annette Kelly
Expected attendee: libraries and other cultural institutions (10)
Questions:
- identification of additional use cases (regularly operating, research, planed) (additional form) navigation…
- benefits expected from GIS in libraries
- geocoding or geotagging geographical coordinates
- geocoding of the historical maps
- GIS in Europeana
- the need for grid and cloud computing

3. table: Museums and cloud computing
Facilitators: Jernej Porenta , Luka Hribar
Expected attendee: museums and other cultural institutions (10)
Questions:
- identification of additional use cases (regularly operating, research, planed) (additional form) navigation…
- comparison the costs (in house server ITC vis-a-vis cloud computing)
- persistent identifiers
- e-infrastructure expected from NRNs