Geo-referencing of biodiversity data in the LifeWatch Italy infrastructure: updates and perspectives

Stefano De Felici, Alessandro Oggioni, Cristiano Fugazza, Fabio Cianferoni, **Paolo Tagliolato***

* LifeWatch Italy, CNR IREA

Conferenza Annuale di LifeWatch Italia
Roma, 25-27 Giugno 2018

What

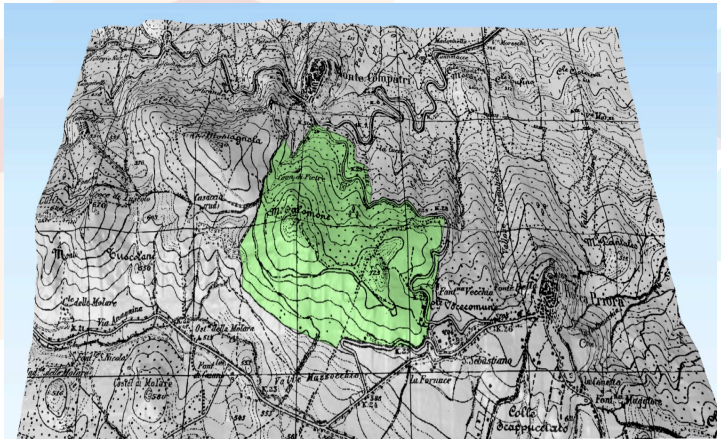
- Construction of a semantically enriched **thesaurus**, to be exploited as the basis of a **toponymical resolution service**, in order to return, given a geographic name, a representative point and its spatial footprint.

Geographic names - georeferencing the data

«We often think of primary species data as being **point records** of plant or animal occurrences but this is **only part of the story**.» (Chapmann, 2005).

Sometimes the samples are directly related to a grid or an area, but also when a record refers a **named place**, it always **describes an area** rather than a true point and that collecting may have occurred anywhere within the described area (Wieczorck et al., 2004).

Lasiommata megera
Monte Salomone
12/06/2004
S.De Felici leg.



Point records of primary specimen records are not really points, but have an error figure associated with them (Chapmann, 2005).

As a result, the outcomes of all the **current methods for georeferencing** primary biodiversity data are **composed of two parts**:

- **a point** that represents the location, and
- **a polygon** that represents the **uncertainty** by which the errors in the transformation process are taken into account.
- **More specific is the description of the locality and smaller could be the polygon.**

General framework – motivation

- Semantic technologies to foster interoperability, discovery, reuse of data and knowledge
- Within ecological sciences: thesauri for authoritative definitions of concepts

General framework – motivation

- Issues highlighted in the past for terms in vocabularies managed with spreadsheets and relational databases that led to the adoption of semantic technologies (cf. Simons, Yu, Cox 2013 "Defining a water quality vocabulary using QUDT and ChEBI"):
  - **Ambiguity**: concepts poorly defined
  - **Inconsistent governance**: same term in multiple vocabularies and relations among them are limited
  - **Lack of modularity**: one discipline needs access, with least effort, to terms from others.
  - **Not interoperable**: **use of local**, non-resolvable **identifiers**, lack of a formal definition, lack of an ontology
- Do the same issues affect geographic names representation management?

Georeference and geographic names: back to the issues

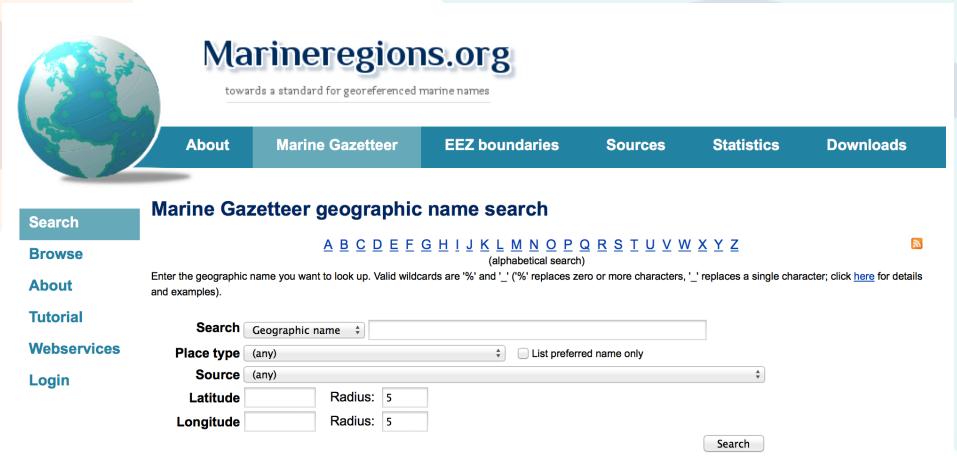
WFS solution

- **Ambiguity**: same geography, different “places” (e.g. Sicily – the Region vs Sicily, the island)
- **Inconsistent governance**. E.g. different WFS define the same geographic features. How to relate them?
- To favour **modularity** (e.g. access to features defined for different disciplines) a solution can be represented by national geoportals but…
- They can use **non persistent identifier** for features (e.g. features ids changing with each request: not possible to use wfs getFeature by id requests as URIs)

Georeference and geographic names: back to the issues

- **For governance, persistent identification**.

Notable example of a strategy for a possible solution : marineregions.org **Marine Gazetteer**. Geographic names have unique MRGID – central gazetteer (rest services). Associated WFS with attribute MRGID.



Lack of a URI for a feature with MRGID.

Proposal, experimentation in progress: semantic geographic features

- Several ontologies to describe geographic features, e.g.
  - skos (Prominent example of skos-gazetteer: gebco features in NVS C19)
  - sweet ontology (realms)
  - geoLink (defines feature types, cf. http://schema.geolink.org/1.0/voc/gebco/featuretype)
- Our choice/proposal (**please, discuss it!**): geonames ontology. Prominent example of gazetteer:
  - geonames.org, rdf downloadable + linked data; queries via REST services, no sparql endpoint)
  - Past experience reported (German federal environmental agency: gein® Gazetteer)
- Work LifeWatch Italia: IGM toponyms (currently served as WFS by National Geoportal) into geonames ontology. (Tagliolato et al. 2017)

Geonames: why?

- Well suited to toponyms
- Mappings (equivalentClasses) to other ontologies (linkedgeodata, geovocab, mindswap geo)
- Hierarchies/relations among geographic features (parent feature, parent country, nearby features)
- Multilingual (trivial: @) and alternate naming support (gn:alternateName sub property of skos:altLabel)

Geographic features as semantic resources: why?

- Use case: historical (and not only) observations located within a “place name” with no coordinates
  - Natural representation with toponyms
- Use case: merging synomyms (owl:sameAs or alternateNames)
- Use case: different samplings of the same feature of interest (e.g. we are studying *this lake*)

IGM Italian toponyms into geonames ontology

- Existing, authoritative sources must be preserved and leveraged.
- the official IGM Italian Toponyms available as WFS from the Italian National Geoportal
- Original source
  - resolution: 1:25.000
  - Well apted to reverse geocoding of historical collections (old geographical names are preserved)

IGM Italian toponyms into geonames ontology

- IGM It toponyms consist in 716.707 points with attributes following the Feature Attribute Coding Catalogue (FACC)
- Administrative inclusion of the toponymes are attributed to points (region, province, city)

Tabella degli attributi - Toponimi :: Totale degli elementi: 11, filtrati: 11, selezionati: 0

toponimo	secondo_nome	tipo	oggetto_toponimo	testo	tavoletta	edizione	data	codice_istat	comune	provincia	regione	cod_comune	cod_pro
SELLA DI LEONE...		NATURALI	PASSO / VALICO	GRANDI	13934XE	1	1955	12057033	LEONESSA	RIETI	LAZIO	033	057
M. TERMINILLO		NATURALI	MONTE / CIMA / CORNO	GRANDI	13934XE	1	1955	12057033	LEONESSA	RIETI	LAZIO	033	057
LE SCANGIVE		ALTRO	AREA GEOGRAFICA	GRANDI	13934XE	1	1955	12057033	LEONESSA	RIETI	LAZIO	033	057
IACCIO CRUDELE		ALTRO	AREA GEOGRAFICA	GRANDI	13934XE	1	1955	12057033	LEONESSA	RIETI	LAZIO	033	057
F.TE CAPO SCURA		IDROGRAFIA	FONTANA	PICCOLI	13934XE	1	1955	12057057	POSTA	RIETI	LAZIO	057	057
M. IL BRECCIARO		NATURALI	MONTE / CIMA / CORNO	GRANDI	13934XE	1	1955	12057057	POSTA	RIETI	LAZIO	057	057
PIAN DI SCURA		ALTRO	AREA GEOGRAFICA	GRANDI	13934XE	1	1955	12057057	POSTA	RIETI	LAZIO	057	057
PRATO CRISTOF...		ALTRO	AREA GEOGRAFICA	GRANDI	13934XE	1	1955	12057057	POSTA	RIETI	LAZIO	057	057
SELLA IACCI		NATURALI	PASSO / VALICO	GRANDI	13934XE	1	1955	12057057	POSTA	RIETI	LAZIO	057	057
VALLONE		ALTRO	AREA GEOGRAFICA	GRANDI	13934XE	1	1955	12057057	POSTA	RIETI	LAZIO	057	057
RIF.O		INSEDIAMENTI	CASE ISOLATE	PICCOLI	13934XE	1	1955	12057057	POSTA	RIETI	LAZIO	057	057

- 114 categories based on FACC.

IGM Italian toponyms into geonames ontology

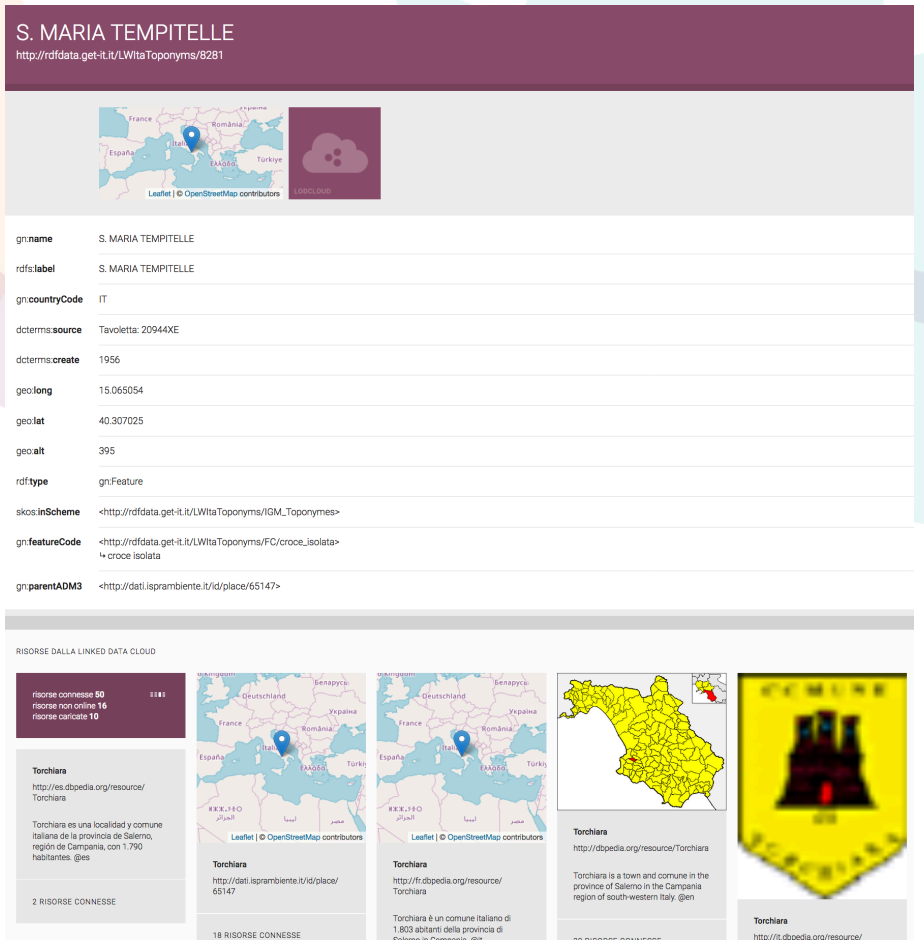
- Work done:
  - mapping of 114 IGM toponymes *categories* to geonames *featureCodes/featureClasses*
    - Issue: despite that FeatureClass and FeatureCodes are an evolution of FACC, currently they scarcely intersect and no official mapping is provided.

argine	H.BNK	bank(s)	an elevation, typically located o	C<D	narrowMatch
bacino montano (lago) (vasca di colmata)	H.RSV	reservoir(s)	an artificial pond or lake	C>D	broadMatch
banchi / massicci rocciosi + scoglio	T.RK	rock	a conspicuous, isolated rocky m	C=D	exactMatch
bastione	S.CSTL	castle	a large fortified building or set o	C>D	broadMatch
bonifica	L.BSND	drainage basin	an area drained by a stream	C=D	exactMatch
bosco	V.FRST	forest(s)	an area dominated by tree vege	C=D	exactMatch
burrone	H.RVN	ravine(s)	a small, narrow, deep, steep-sid	C=D	exactMatch

    - Results:
      - 27 categories are not mapped to geonames (153k points)
      - 7 categories have multiple correspondent geonames codes/classes
      - 9 categories mapped to 4 featureCodes
  - Mapping of 720.000 toponymes to RDF representation

IGM Italian toponyms into geonames ontology

- Work done (2017):
  - complete RDF representation available within a triple store.
  - Resources available through:
    - SPARQL endpoint
    - Linked open data access
  - Resources are linked to ISPRA and ISTAT Linked Open Data
  - See examples within the **deployed portrayal service** at http://lodview.get-it.it/



IGM Italian toponyms into geonames ontology

- Upgrades (2018):
  - Till now (original source): toponym -> point
  - But the implied knowledge of a place name used for georeferencing observations comprises (cf. Chapman & Wieczorek, 2006):
    - **Extent** (measure of the size of the area within which collecting observations occurred)
    - **Footprint** (spatial representation of a feature as an area)
    - **Uncertainty** (measure of the incompleteness of one's knowledge about an unknown quantity)
  - How to construct them?
    - → using the feature codes (categories) available for each name in the original source

IGM Italian toponyms into geonames ontology

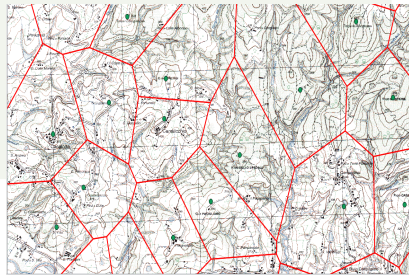
- Upgrades (2018):
  - feature codes → feature types (points, lines, polygons)

POINT	LINE	POLYGON
ABBEVERATOIO / ABBEVERATOIO CON FONTANA	ACQUEDOTTO	AEROPORTO
AEROMOTORE	ARGINE	AREA ALLEVAMENTO BESTIAME
ALPE / CASA DI ALTA MONTAGNA	CANALETTO DI IRRIGAZIONE MONTANA	AREA ARCHEOLOGICA
ANTENNA PER TELECOMUNICAZIONI	CARRARECCIA B2	AREA GEOGRAFICA
BASTIONE	FERROVIA	BACINO MONTANO (LAGO) (VASCA DI COLMATA)
BOA	FESSURA / CREPACCIO	BANCHI / MASSICCI ROCCIOSI
BURRONE	FIUME / TORRENTE	BONIFICA

  - Feature types → tentative standardization of extent, footprint and uncertainty for each toponyms, by means of declared procedures

IGM Italian toponyms into geonames ontology

Feature type	footprint	Extent/uncertainty	Additional source
Point (e.g. wells)	Circle (r=extent)	250m	-
Line (e.g. rivers)	Buffer along line	50m-100m	Elementi idrici (WFS national geoportal)
Polygon (defined boundaries) (e.g. lake, towns)	Area defined by boundaries		Administrative divisions, protected areas, etc.
Polygon (undefined boundaries – e.g. mountains, valleys)	Tile in a **Voronoi tessellation** of points from the same feature class	Distance of the centroid from the farthest point on the boundary	-



Geographic features as semantic resources: perspectives

- update the current semantic assets with the prepared information
- enable indices (spatial and textual) on the triple store
- furtherly test querying capabilities with the indices
- activate reasoning capabilities (at least) for implicit predicates
- test the scenario of a geocoding service.