

Ornithology Based on Linking Bird Observations with Weather Data

Mikko Koho^{1,2,3(✉)}, Eero Hyvönen^{2,3}, and Aleksi Lehikoinen⁴

¹ CSC - IT Center for Science Ltd., Espoo, Finland
`mikko.koho@csc.fi`

<http://www.csc.fi/english>

² Semantic Computing Research Group (SeCo), Aalto University, Espoo, Finland

³ Semantic Computing Research Group (SeCo),
University of Helsinki, Helsinki, Finland

`eero.hyvonen@aalto.fi`

<http://www.seco.tkk.fi/>

⁴ The Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland

`aleksi.lehikoinen@helsinki.fi`

<http://www.luomus.fi/en/>

Abstract. This paper presents the first results of a use case of Linked Data for eScience, where 0.5 million rows of bird migration observations over 30 years time span are linked with 0.1 million rows of related weather observations and a bird species ontology. Using the enriched linked data, biology researchers at the Finnish Museum of Natural History will be able to investigate temporal changes in bird biodiversity and how weather conditions affect bird migration. To support data exploration, the data is published in a SPARQL endpoint service using the RDF Data Cube model, on which semantic search and visualization tools are built.

1 Serving Ornithologists with Linked Data

Long-term standardized nature observations provide crucial data for determining population trends in biodiversity management. For example, from monitoring data about bird migration over many years, it is possible to determine trends [13, 20] in population size, and annual breeding success [12]. By combining observational data with related datasets, such as weather or pollution data, it is possible to investigate how biodiversity is affected by related phenomena in nature.

This paper presents a use case of adding value to a bird observation dataset by related weather data and a species ontology using the Linked Data [16] approach. We first describe the data, metadata schemas, and ontologies used. After this, publication of the data on a SPARQL endpoint is discussed. Based on the SPARQL endpoint's API, tools and visualizations for researching and exploring the data are being created for the ornithologists, bird watchers and researchers alike, to use—two examples of this are explained in Sect. 3. In conclusion, contributions of our work and related research are discussed.

2 Data, Metadata, and Ontologies

Data. The bird observation data comes from the Hanko Bird Observatory “Halias” located in Hanko, Finland. The data has been gathered systematically since 1979 and is actively used in scientific research¹.

Nature observation datasets can be divided into two main classes:

1. **Systematic observational datasets.** The methods used in making nature observations are standardized and every species in the research scope is counted. Straightforward to do inference e.g. about spatio-temporal occurrence of species.
2. **Open public datasets.** People are free to report nature observations they have made. Methods used in making observations vary and only some portion of observed species may be reported. Any inference done requires knowledge about the way people report observations and results can be misleading.

The Halias dataset is a systematic observational dataset consisting of daily bird counts. All migrating birds are counted daily in a standardized way for four hours starting from sunrise. In addition, local as well as migrating birds outside the standardized observation time are counted for each day. Every observed bird is counted and daily counts are given for each date and species combination.

The weather data comes from a Finnish Meteorological Institute’s weather observation station in Russarö, situated less than 6 km away from Halias. The weather observation data consists of measurements of temperature, relative humidity, wind direction (10 min average), wind speed (10 min average), cloud coverage and sea level air pressure, collected every 3 h. Also a total rain amount is given for each day.

Metadata Schema. We have created a data schema for representing the daily bird observations based on the RDF Data Cube Vocabulary² of W3C. The schema heavily re-uses—for interoperability reasons—other vocabularies where possible, such as Darwin Core³, TaxMeOn [21], and data.gov.uk Time Intervals⁴.

Observed daily weather variables are aggregated based on needs of this use case. Weather observations that correspond to the standardized observation times of each day are of particular interest. They make it possible to match the weather data and bird migration counts on a shorter time span than the whole day and thus giving better accuracy for finding the connections between them. We calculate average values of the weather variables for the 4-hour standardized observation time. The average values are calculated assuming a linear change in a weather variable’s values between the 3-hour observations. The aggregated values of weather variables are then linked directly to the bird observations.

¹ See <http://www.tringa.fi/web/lintuasemat/hangon-lintuasema/julkaisuluettelo.html> for a list of research publications related to observations at Halias.

² <http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/>

³ <http://rs.tdwg.org/dwc/>

⁴ <http://datahub.io/dataset/data-gov-uk-time-intervals>

Wind observations are represented using an ontology of all observed wind speed and direction combinations. The observed winds are then linked with the daily bird observation instances. The schema is documented in detail online⁵.

We envision, that by using a standard data model for observations, our forthcoming tools for exploring and visualizing observational data could be applied not only to birds but to other observations, too, and conversely, tools developed by others for the Data Cube model, could be re-used in our case study more easily.

Ontologies. The data is linked with the AVIO Ontology of the Birds of the World [22], based on data from the taxonomic database⁶ of the Finnish Museum of Natural History.

For our Halias case study, the original AVIO was refined and extended as follows:

1. Scope of the species ontology was reduced to species observed in Finland. This is due to the need for generating species name abbreviations by certain rules which do not apply globally.
2. Species name abbreviations were added as they are used in the primary data.
3. Species characteristics based on the characteristics system used in Nature Gate service⁷ were added.
4. Species conservation statuses were added in both Finnish and EU scope.
5. Rarity classes were added. These are subjective views of rarity of species.

The species characteristics ontology classifies birds in terms of four facet categories: (1) Date and nesting habitat, (2) Coloring and markings, (3) Shape and size, and (4) Behavior. These categories are then divided further into subcategories.

Linked Data Publication. The datasets were combined into a single linked dataset. This linked dataset and the used ontologies compose the whole linked data publication, which consists of 12,315,709 triples.

The data was published as a SPARQL service using the Linked Data Finland (LDF) [9] platform⁸. Based on metadata about the dataset and its graphs, this platform automatically generates APIs and content negotiation services for machines to use, and a homepage for human users with ready to use services, such as Linked Data browsing and editing the data.

The bird observation data is not openly available, but there are ongoing discussions about making the data more openly accessible. Access to the published data is currently restricted to the association that owns the data and collaborating research projects. Nevertheless, the data can be requested for research purposes without any costs.

⁵ <http://www.essepuntato.it/lode/http://ldf.fi/schema/halias/>

⁶ <http://taxon.luomus.fi/>

⁷ <http://naturegate.fi/>

⁸ <http://www.ldf.fi/>

Some structural problems in the original bird observation data were noticed when the transformed linked data was first visualized. There were duplicate species and date combinations with different bird counts, when there should never be more than one of these combinations and the URIs of the observations were created using these combinations. This anomaly was taken into account in the transformation process by creating different URIs for each of the duplicate combinations.

After this, more validation of the data was done in the transformation process and we discovered some other problems of lesser impact, which did not hamper the linked data publication or visualization.

3 Visualizing the Data to Solve Problems

There is a strong connection [1,2,15] between the bird migration magnitude and weather variables of a day in the spring and autumn bird migration periods. Finding these connections from the data however is not a trivial task. Our approach to this problem is to provide data visualization tools based on linked data. Such tools let the ornithologists to explore, find, and visualize relationships between variables in bird and weather observation data. Good visualizations would not only show interesting correlations and patterns in data, but also raise up new questions for further data exploration and research.

Visualizing the data is flexible on top of a SPARQL endpoint with commonly used tools. For example, Sgvizler⁹ [19] is a tool that is easy to use for combining SPARQL queries with Google Charts¹⁰ graphics. We next exemplify these possibilities by two visualizations combining bird and weather observations.

Our first visualization originates from the research problem of a collaborating biologist, where he wished to investigate, how local wind conditions affect the migration of different species, or groups of species, during spring and autumn migration. To solve the problem, the visualization of Fig.1 was created. Here the idea is to use a Circular heat chart¹¹. The diagram shows the relationship between wind conditions and observed crane (*Grus grus*) migration in autumn months (September, October, and November). For each observed wind speed and direction combination, the number of migrating cranes during days in which the wind combination is observed, is added up. These sums are then normalized by the number of days with the wind combination to account for the biased occurrence of different wind conditions. Directions in the diagram correspond to wind directions using standard English abbreviations, and distance from the diagram center represents wind speed in meters per second. Migration magnitude increases from gray (no migration) to green. The color scale is normalized so that the deepest green shade corresponds to maximum migration magnitude observed.

During autumn, cranes migrate from breeding areas in Finland to southern wintering areas. As large birds, they are expected to favor tail winds as they

⁹ <http://dev.data2000.no/sgvizler/>

¹⁰ <https://developers.google.com/chart/>

¹¹ <http://prcweb.co.uk/lab/circularheat/>

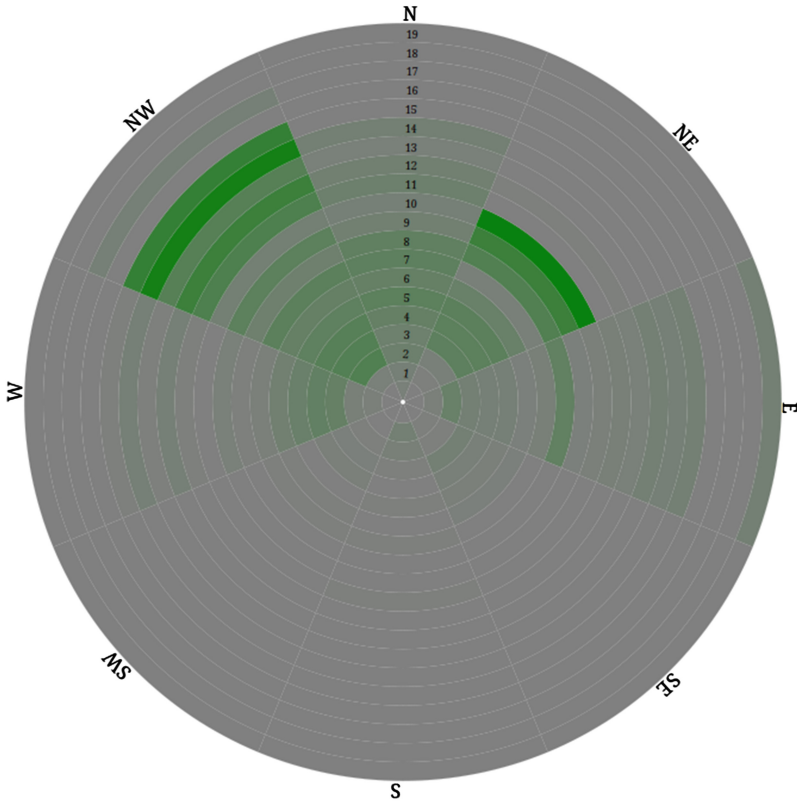


Fig. 1. Diagram of crane migration magnitude on days with different wind conditions in autumn.

migrate southward from Hanko over the Baltic Sea. The example visualization supports this hypothesis.

The visualization is based on a SPARQL query. The graphics were implemented using D3¹² JavaScript library and AngularJS¹³ JavaScript framework. The SPARQL query used takes about five seconds to retrieve the results from the LDF platform. Here is the SPARQL query for this visualization:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX h: <http://ldf.fi/schema/halias/>

SELECT ?order ?speed ((SUM(?cnt)/?days) AS ?normalized)
WHERE {
    ?observation h:observedSpecies ?species .
    ?species rdfs:label "Grus grus".
    ?observation h:season h:autumn .
```

¹² <http://d3js.org/>

¹³ <http://angularjs.org/>

```

?observation h:countMigration ?cnt .
?observation h:windDay ?wind .
?wind h:windSpeed ?speed .
?wind h:windDirection ?dir .
?dir h:order ?order .
{
  SELECT ?wind (COUNT(DISTINCT ?date) AS ?days)
  WHERE {
    ?observation h:refTime ?date .
    ?observation h:windDay ?wind .
    ?observation h:season h:autumn .
  }
  GROUP BY ?wind
}
}
GROUP BY ?order ?speed ?days
ORDER BY ?speed

```

The research hypothesis behind our second visualization is that air pressure affects the migration of very large birds. However, it is unclear how much it affects, if at all. To solve the problem, the characteristics ontology of bird species is needed telling the sizes of different species, in addition to bird observation and weather data. Figure 2 plots as an example the migration counts of very large birds on Y-axis. The average air pressure of the days in which the migration was observed is represented in the X-axis. The migration counts have been normalized to give the average migration count for each day with a certain average air pressure.



Fig. 2. Average migration count of very large birds in relation to air pressure [hPa].

This visualization was implemented using Sgvizler. It takes about 8 seconds to retrieve the data from the LDF platform using this SPARQL query:

```

PREFIX h: <http://ldf.fi/schema/halias/>
PREFIX bc: <http://ldf.fi/halias/bird-characteristics/>

SELECT ?pressure ((SUM(?cnt)/?days) AS ?normalized)
WHERE {
    ?observation h:observedSpecies ?species .
    ?species h:hasCharacteristic bc:valtava .
    ?observation h:airPressure ?pressure .
    ?observation h:countMigration ?cnt .
}
{
SELECT ?pressure (COUNT(DISTINCT ?date) AS ?days)
WHERE {
?observation h:refTime ?date .
?observation h:airPressure ?pressure .
}
GROUP BY ?pressure
}
}
GROUP BY ?pressure ?days
ORDER BY ?pressure

```

Based on the graph, it may be inferred that very large birds migrate mostly during high pressure when there are low chances of rain. As the normal air pressure is around 1013, this seems to be in line with the original hypothesis. Also it was observed that when visualizing the migration magnitude of different bird sizes, smaller birds had higher migration magnitudes at lower air pressures.

Because very low or very high air pressures are rare, a large variance in the values far away from the normal air pressure can be seen in the graph.

4 Discussion

4.1 Contributions

This paper suggests that following advantages may be gained by publishing a statistical dataset using the RDF Data Cube Vocabulary, and by enriching the data by linking it with related datasets:

1. **Better understanding of the primary data.** The weather variables play a key role in the migration magnitude of a single day. Thus by introducing the weather context of the time of observation, a better understanding about the bird observation data can be achieved.
2. **Interoperability.** Efficient use of nature observation data requires standard ways to combine datasets. This can be achieved with by publishing datasets as linked data [16], which makes it also possible to use common tools for data handling and processing [17].

3. **Visualization tools.** It is possible to visualize the data with existing RDF Data Cube visualization tools [14,17] and in the future with new and better tools.
4. **Linking to taxonomic concepts.** By linking to a taxon ontology we are able to express certain taxonomic concepts of the species instead of representing them only by name, which then would need work from an expert to correctly interpret [4,18]. One way to express taxonomic concepts and changes biological conception is to use the TaxMeOn meta-ontology [21].
5. **Data validation.** Structural errors in original data, such as wrong data types or cardinality errors, may cause problems when converting to RDF and thus the conversion may provide information about errors in the original data and help improve data quality.

Performance of SPARQL queries may be an issue in visualizing large linked datasets. The linked dataset used in this research is already large enough to make the data retrieval last several seconds.

4.2 Related Work

There are bird observation analysis systems around, such as eBird¹⁴ used by, e.g., the Audubon Society, that provide visualizations of observations, such as range and point maps and yearly bar charts. The different metrics of observations can be graphed along a timeline and statistics of one species contrasted with others. Furthermore, data mining tools have been applied to observational databases in order to analyze and discover phenomena that take place in nature [7]. Our approach extends these ideas to the Semantic Web, especially to using enriched semantic data based on data linking and ontologies.

The AVIO ontology and species characteristics ontology have been used in our earlier work, the BirdWatch mobile tool for ornithologists [8].

Lots of Linked Data (LD) platforms have emerged on the Web since the publication of the four Linked Data publication principles and the 5-star model¹⁵ [6]. For example, in Life Sciences alone there are LinkedLifeData¹⁶, NeuroCommons¹⁷, Chem2Bio2RDF¹⁸, HCLSIG/LODD¹⁹, BioLOD²⁰, and Bio2RDF²¹. The Linked Data Finland service used in our case study is based on a “7-star model” with additional services supporting dataset schema documentation, data validation, and data curation [9].

Work in Linked Data visualization [3,11] falls in three categories. One can (1) visualize the structure of data, e.g., ontologies in an ontology editor, (2) present

¹⁴ <http://ebird.org/>

¹⁵ <http://www.w3.org/DesignIssues/LinkedData.html>

¹⁶ <http://linkedlifedata.com/>

¹⁷ <http://neurocommons.org/>

¹⁸ <http://chem2bio2rdf.wikispaces.com/>

¹⁹ <http://www.w3.org/wiki/HCLSIG/LODD>

²⁰ <http://biolod.org/>

²¹ <http://bio2rdf.org/>

computed analysis results, e.g., statistics, or (3) use various graphical means for illustrating phenomena, e.g., present data on maps, timelines or using other kind of graphics. Our work is related to the latter two categories, where e.g. business graphics and visualizations been widely and routinely used for data analysis and exploration.

The two visualizations presented in this paper are examples of how Linked Data can be exploited to address research questions in an application domain. Based on the experiments, the SPARQL standard was deemed quite flexible in formulating fairly complex queries, and using visualization frameworks on top of query results was fairly straightforward. Based on our hands-on experiments, data visualization seems to be a promising application dimension of Linked Data. Visualizations, such as our examples, are focused on solving particular problems, but at the same time may well generalize to other use cases by changing the variables involved, such as the species, species group criterion, time frame, or a weather characteristic.

Interest in publishing statistical data as linked data has been growing [5, 10] in the past few years and has been studied in many publications. Hausenblas et al. [5] have proposed publishing statistical datasets as linked data using the Statistical Core Vocabulary (SCOVO) and a modelling framework to support this.

Salas et al. [17] present tools OLAP2DataCube and CSV2DataCube for transforming statistical data to RDF Data Cube Vocabulary datasets. Using general purpose link discovery tools for a transformed dataset is shown through an example. Also presented is an extension to OntoWiki for visualizing RDF Data Cube datasets called CubeViz.

Mutlu et al. [14] present a visualization wizard of linked research data that automatically suggests suitable visualizations based on the data and a semantic configuration. Kämpgen and Harth [10] discuss using open-source OLAP systems to visualize statistical Linked Data.

4.3 Future Research

The visualizations have been implemented with this particular use case in mind. We should study if a more general RDF Data Cube visualization approach would suit the needs of this use case. A big challenge with a more general approach comes from the user interface, which should be kept simple and easy to use.

Some existing open source RDF Data Cube visualization services were evaluated when we decided to use the data cube approach, and they were not up to the task at that point. As there seems to have been development in this area, a new evaluation of them would be useful.

This combined data of bird observation counts and weather variables seems interesting for new research in both computer science and biology.

Acknowledgements. This work has been supported by the Ministry of Education and Culture, CSC - IT Center for Science Ltd., National Research Data Initiative

(TTA)(<http://www.tdata.fi/en/>), and the Linked Data Finland project (<http://www.seco.tkk.fi/projects/ldf/>) funded by the Finnish Funding Agency for Innovation (Tekes) of 20 partners.

Ornithological society Tringa(<http://www.tringa.fi/web/>) provided the data used in this research. Hundreds of volunteers at Halias have helped by gathering the data. Miika Alonen from Aalto University's Semantic Computing Research Group (SeCo) gave support in visualizing the data. Jouni Tuominen from SeCo contributed to developing the visualization web service.

References

1. Able, K.P.: The role of weather variables and flight direction in determining the magnitude of nocturnal bird migration. *Ecology* **54**(5), 1031–1041 (1973). <http://www.jstor.org/stable/1935569>
2. Alerstam, T.: Optimal bird migration revisited. *J. Ornithol.* **152**(1), 5–23 (2011). <http://dx.doi.org/10.1007/s10336-011-0694-1>
3. Dadzie, A.S., Rowe, M.: Approaches to visualising linked data: a survey. *Semant. Web* **2**(2), 89–124 (2011)
4. Franz, N.M.: Biological taxonomy and ontology development: scope and limitations. *Biodivers. Inform.* **7**(1), 45–66 (2011)
5. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., Ayers, D.: SCOVO: using statistics on the web of data. In: Aroyo, L., et al. (eds.) *ESWC 2009*. LNCS, vol. 5554, pp. 708–722. Springer, Heidelberg (2009)
6. Heath, T., Bizer, C.: *Linked Data: Evolving the Web into a Global Data Space*, 1st edn. Morgan & Claypool, Palo Alto (2011). <http://linkeddatatoolkit.com/editions/1.0/>
7. Hochachka, W., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., Kelling, S.: Data-mining discovery of pattern and process in ecological systems. *Wildl. Manag.* **71**(7), 2427–2437 (2007)
8. Hyvönen, E., Alonen, M., Koho, M., Tuominen, J.: Birdwatch—supporting citizen scientists for better linked data quality for biodiversity management. In: *Proceedings of the First International Workshop on Semantics for Biodiversity (S4BioDiv)*, *ESWC 2013*. CEUR Workshop Proceedings, vol. 979, May 2013. <http://www.ceur-ws.org/Vol-979>
9. Hyvönen, E., Alonen, M., Tuominen, J., Mäkelä, E.: Linked data finland: towards a 7-star service platform for linked datasets. In: *The First Annual KnowEscape Conference - KnowEscape 2013*, November 2013. <http://www.seco.tkk.fi/publications/2013/hyvonen-et-al-ldf-2013.pdf>
10. Kämpgen, B., Harth, A.: Transforming statistical linked data for use in olap systems. In: *Proceedings of the 7th International Conference on Semantic Systems*, pp. 33–40. ACM (2011)
11. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E.: Ontology visualization methods—a survey. *ACM Comput. Surv. (CSUR)* **39**(4), 10 (2007)
12. Kjellén, N.: Annual variation in numbers, age and sex ratios among migrating raptors at falsterbo, sweden from 1986–1995. *J. für Ornithologie* **139**(2), 157–171 (1998)
13. Lehikoinen, A., et al.: Lintukantojen kehitys hangon lintuaseman aineiston mukaan 1979–2007. *Tringa* **35**, 313–321 (2008)

14. Mutlu, B., Hoefler, P., Sabol, V., Tschinkel, G., Granitzer, M.: Automated visualization support for linked research data. In: Proceedings of the I-SEMANTICS 2013 Posters & Demonstrations Track, p. 40 (2013)
15. Nisbet, I.C.T., Drury Jr., W.H.: Short-term effects of weather on bird migration: a field study using multivariate statistics. *Anim. Behav.* **16**(4), 496–530 (1968)
16. Reichman, O., Jones, M.B., Schildhauer, M.P.: Challenges and opportunities of open data in ecology. *Science(Washington)* **331**(6018), 703–705 (2011)
17. Salas, P.E.R., Auer, S., Breitman, K.K., Casanova, M.A., Martin, M.: Publishing statistical data on the web. *Int. J. Semant. Comput.* **6**(4), 373–388 (2012)
18. Schulz, S., Stenzhorn, H., Boeker, M.: The ontology of biological taxa. *Bioinformatics* **24**(13), 313–321 (2008)
19. Skaeveland, M.: Sgvizler: a JavaScript wrapper for easy visualization of sparql result sets. In: Proceedings of the ESWC 2012. Springer-Verlag (2012)
20. Svensson, S.E.: Efficiency of two methods for monitoring bird population levels: breeding bird censuses contra counts of migrating birds. *Oikos* **30**, 373–386 (1978)
21. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological names and taxonomies on the semantic web – managing the change in scientific conception. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part II. LNCS, vol. 6644, pp. 255–269. Springer, Heidelberg (2011)
22. Tuominen, J., Laurenne, N., Koho, M., Hyvönen, E.: The birds of the world ontology AVIO. In: Cimiano, P., Fernández, M., Lopez, V., Schlobach, S., Völker, J. (eds.) ESWC 2013. LNCS, vol. 7955, pp. 300–301. Springer, Heidelberg (2013)