A Measure for Relevance of User Needs, Observational Requirements, and Earth Observations Data

Hans-Peter Plag[1] & Jay Pearlman[2] & Francoise Pearlman[2]
1) Nevada Bureau of Mines and Geology, University of Nevada, Reno
Mail Stop 178, Reno, Nevada 89557, USA, hpplag@unr.edu,
2) J&F Enterprise

April 17, 2012

DRAFT VERSION 0.2 — DO NOT CITE

Abstract

The User Requirements Registry (URR) of the Global Earth Observation System of Systems (GEOSS) is an integral part of the GEOSS Common Infrastructure (GCI) used to publish user- related information and to provide analysis tools that allow answering questions related to a user- driven design and functionality of GEOSS. The URR collects user-related information, such as the user types, their applications and activities, the requirements of the applications in terms of observations and other products, and the needs in terms of research, infrastructure, technology, and capacity building that would enable or improve applications. This information is collected in standard formats and nomenclature across disciplines and Societal Benefit Areas (SBAs). Interconnectivity of user types, applications, requirements and needs is captured with a novel link concept, and this information enables prioritization of applications, requirements and needs, gap analyses, and the determination of the relevance of a given product.

Prioritization uses the interconnectivity captured in the links to determine the relevance of observational requirements, applications, and needs in infrastructure, research, technology, and capacity building. The interconnectivity between all of these entities can be used to construct answers to the generic question "How relevant is entry A?" It is of value to apply the concept of relevance not only to requirements. For example, answers to questions like "How relevant is this research need?" or "How important is this infrastructure?" could help prioritize research, development and implementation efforts. For datasets and products, the core question is "What is the societal relevance and potential impact of this dataset?"

It will be important to achieve a GEO-wide consensus on the quantitative metrics for measuring relevance. Different options for the metrics for measuring the relevance of requirements, applications, or other needs, as well as the relevance of a dataset are being discussed and assessed. For a given URR entry of a requirement, application, or need, a simple measure could consist of the number of other entries depending directly or indirectly on this particular entry. A more comprehensive measure would take into account the relevance of the linked dependent entries, especially the applications, and the strength of the links between entries. Further factors could account for the societal value of the links in the measure. A measure of relevance can also be deduced from the status of an entry in the mathematical network constituted by the entries and the links between them. Mathematical network theory can be used to support the measure of relevance. Factors such as pervasiveness, uniqueness, difficulty, completeness, and uncertainty are also candidates to be considered. The challenge is to balance quantitative and qualitative metrics in forming a coherent evaluation. These subjects will be addressed at the workshop. For products (observations, derived products, or services), a metric for measuring relevance could be connected to the relevance of the observational requirements this product meets. If a product meets more than one observational requirement, a challenge is in the combination of the relevance of the individual requirements into an overall relevance measure for the products.

1 Introduction

The *Global Earth Observation System of Systems (GEOSS)* developed by the *Group on Earth Observations (GEO)* is intended as a user-driven system of system (GEO, 2005). A challenge in developing a user-driven observing system is the prioritization of user needs, and the determination of the relevance of existing systems and the datasets and products they provide. the *GEOSS Common Infra Structure (GCI)* of GEOSS collects user-related information in the *User Requirement Registry (URR)* (Plag, 2012). Based on the information collected in the URR, it is possible to measure the relevance of an entry in the URR or of an available dataset or product. In order to have credibility, the metrics for measuring relevance will have to be based on a GEO-wide consensus.

Originally, prioritization focused on requirements. However, there is a significant benefit in being able to prioritize the applications, as well as the needs mentioned above. For example, an answer to the question of "How important is this research need?" could guide researchers and support them in soliciting funds for their research. Answering the question "How important is this infrastructure need?" would help to prioritize efforts and secure funding for the implementation of important infrastructure.

This paper discusses the background for a relevance measure and proposes a metric to be used. There are many options for the metric. A simple measure could consist of the number of other entries depending on entry A or dataset B. A more comprehensive measure would take into account the relevance of the linked dependent entries, especially the applications, and the strength of the links between entries. The societal value of the links in the measure could also be used as part of the metrics.

A measure of relevance can also be deduced from the status of an entry in the mathematical network constituted by the entries and the links between them. We discuss to what extent network theory can be used to support a measure of relevance. Factors such as pervasiveness, uniqueness, difficulty, completeness, and uncertainty will be considered.

2 Relevance Related Parameters

The measure of relevance r for the various entries in the URR as well as for datasets and products will be a function of a number of measurable or quantifiable parameters. We express r as

$$r = f(p_1, p_2, ..., p_n), \tag{1}$$

where $p_i,\ i = 1, ..., n$ are quantitative (measurable) and qualitative (but quantifiable) parameters.

Potential candidates for the parameters determining the relevance of an entity A (entry in the URR or dataset outside the URR) are the number and weight of links for which A is the source, the relevance r_i of each target B_i to which A is linked, the relative position in a value chain. Characteristics such as pervasiveness, uniqueness, difficulty, completeness, and uncertainty also could be considered.

Our first task is to identify all parameters that somehow relate to the relevance of an entry in the URR and to assess to what extent they are measurable or quantifiable in an objective way. The next step is then to develop a function f that combines these parameters into a single number, i.e., the relevance r. In the following, we develop a matrix of the parameters and describe their characteristics in terms of quantification (see Table 1).

Comment: Here we need to develop a more complete matrix of potential parameters (see Table to be used in the next Section to define a set of metrics. For datasets, continuity (length of time series) needs to be accounted for. The uniqueness of a dataset also increases the relevance. For a requirement, the frequency of its appearance in various databases indicates relevance. For both, datasets and requirements, user rating can be included as a component of relevance.

3 The Relevance Function

We will develop the relevance function in several steps, starting with a straight-forward part and then moving on to more intricate functions.

Table 1. Parameters that are related to the relevance of a URR entry. Column Q has a (Y)es, if the parameter is quantitative. Column W has a (Y)es if the parameter can be weighted.

Parameter	Q	Rationale	Measurement	W	Comments
No. sources	Y	The relevance of entry A directly relates to the no. of entries depending on A.	Counting the links with A as source.	n/a	Can be determined based on URR contents. Is being used indirectly.
Strength of links	N	The relevance of entry A can depend on the strength of the links with A as the source. The strenght can be converted into a weight.	Link attribute.	n/a	Can be determined based on link attribute. Needs to be converted into a quantitative weight.
Relevance of targets	Y	The relevance of entry A can depend on the relevance of the targets of the links with A as the source.	Relevance of all target entries.	Y	Can be determined starting with those entries that do not appear as a source, and then working backwards.

A rather straight-forward measure is based on the number of links in which entry A is the source or the number of requirements a dataset meets. For an entry A, we can define a local relevance r^{loc} as

$$r^{\mathrm{loc}}(A) = \sum_{i=1}^{L_A} w_i, \tag{2}$$

where L_A is the number of links in which A appears as the source, and $w_i, i = 1, ..., L_A$ are the weights assigned to each link. In the most simple case, we can set all $w_i = 1$. Taking into account the the strength of a link (weak, strong, crucial), we could assign the values 1, 2, or 3 for weak, strong and crucial links, respectively.

The next level of complexity would take into account the relevance of the target. In this case, all relevances would become interdependent and would have to be determined in a recursive way. In this case, publishing a new entry and linking it to one other entry has the potential to change the relevance of a large number of other entries. However, at this level, the relevance would not just be derived from the immediate environment of an entry but reflect its global relevance. A measure for the global relevance r^{glob} is defined here as

$$r^{\mathrm{glob}}(A) = \sum_{i=1}^{L_A} r_i^{\mathrm{glob}} \tag{3}$$

Combining equations (2) and (3), we can define a weighted global relevance by

$$\hat{r}^{\mathrm{glob}}(A) = \sum_{i=1}^{L_A} w_i \cdot \hat{r}_i^{\mathrm{glob}} \tag{4}$$

Based on the global relevance of requirements, we can define the relevance $r(B)$ of a dataset or product B. If B meets the requirements $R_i, i = 1, ..., K_B$, then the relevance of B is defined as

$$r^{\mathrm{glob}}(B) = \sum_{i=1}^{K_B} \hat{r}_i^{\mathrm{glob}}. \tag{5}$$

Since there is no predefined strength of the match between a dataset and requirement entry, we cannot define a weighted relevance here.

In summary, we propose to use the following two measures for relevance of URR entries and external datasets:

For URR entries:

$$\hat{r}^{\text{glob}}(A) = \sum_{i=1}^{L_A} w_i \cdot \hat{r}_i^{\text{glob}}, \tag{6}$$

L_A: number of entries E_i that are targets in the links with entry A being the source;
w_i: weight of the link between A and E_i;
r_i^{glob}: global relevance of E_i.
For external datasets or products:

$$r^{\text{glob}}(B) = \sum_{i=1}^{K_B} \hat{r}_i^{\text{glob}}, \tag{7}$$

K_B: number of requirement entries R_i that are met by dataset or product B;
r_i^{glob}: global relevance of requirements R_i.

4 Algorithms

Tbw

Acronyms

GCI GEOSS Common Infra Structure

GEO Group on Earth Observations

GEOSS Global Earth Observation System of Systems

URR User Requirement Registry

References

GEO, 2005. The Global Earth Observing System of Systems (GEOSS) - 10-Year Implementation Plan, Available at http://earthobservations.org.

Plag, H.-P., ed., 2012. *The GEOSS User Requirement Registry (URR): A Component of the GEOSS Common Infrastructure in Service of Work Plan Implementation and a User-Driven GEOSS*, Group on Earth Observations.