Discovery in Spatial Data Infrastructures (SDIs) is an issue since SDIs made their first appearance in the late 1990s. Naturally, discovery is also a challenge for INSPIRE. In a recent effort, we have, in cooperation with the EEA, extended the General Environmental Multi-lingual Thesaurus (GEMET) to include the INSPIRE spatial data themes. In this abstract, we present first ideas on how the extended GEMET can be used to support the creation of metadata and discovery of data within INSPIRE.

String-matching search terms with metadata keywords is a common mechanism used for discovering datasets and services in SDIs. Anyone who ever used this mechanism is familiar with the following problem: A provider used a certain keyword in the metadata to describe a resource. Unfortunately, the requestor who tries to discover the resource uses a different spelling, a synonym, or a broader or narrower term, and thus does not find the resource. Also, the resource metadata might be in a language that the requestor is unfamiliar with - a major challenge for INSPIRE, which has to deal with the enormous amount of the EU’s 23 official languages.

The INSPIRE Implementing Rule for Metadata (MD IR) addresses these issues for spatial datasets and spatial dataset series through the use of a controlled vocabulary: It requires that “at least one keyword shall be provided from the General Environmental Multi-lingual Thesaurus (GEMET) describing the relevant spatial data theme”. In cooperation with the European Environment Agency (EEA) we have therefore extended GEMET with these keywords and their definitions in all 23 languages. GEMET however is more than a selection of keywords and their translations: It presents a taxonomy of concepts related to each other by broader term, narrower term and related term relationships. In order to make better use of this structure, we have mapped the INSPIRE themes onto related GEMET concepts. In the following, we describe a number of ideas on how this setting can support both the creation of metadata and search for datasets in the context of INSPIRE.

In the creation of metadata, the extended GEMET can help providers in selecting GEMET keywords for annotating their data. After selecting an INSPIRE theme in the language of his choice, a provider can choose further keywords among all GEMET concepts that are related to this theme using one of the following approaches:

- **Using metadata statistics**: The data provider is presented with a cloud representation of GEMET concepts linked to INSPIRE themes. The size of each concept depends on how often it has been used as a keyword to annotate a dataset. This representation is independent of the GEMET classification. It will support a natural selection of terms and result in a set of frequently used keywords (Figure 1).

---

1 European Commission – Joint Research Centre, Institute for Environment & Sustainability, Via Fermi 2749, 21027 Ispra, Italy, {nicole.ostlaender|michael.lutz}@jrc.it
Using the GEMET structure: The data provider is presented with a tree representation of the GEMET taxonomy of concepts linked to the INSPIRE theme. He may search all sub-concepts and decide for one or more additional keywords among narrower terms. This approach is independent from how often a keyword has been used before (Figure 2).

Mixed approach: The data provider is presented with a tree (or graph) representation, where each keyword has a size that depends on how often it has been used in the metadata. This approach combines the benefits of both the statistical and the structured approach. However, while the first two approaches are rather well established, the suggested combination of both is novel and might therefore be more difficult to apply by inexperienced users (Figure 3).

In discovery the extended GEMET can help requestors in selecting search term(s). A requestor chooses his search term(s) based on one of the following approaches, which are based on the same principles as the annotation of data:
• **Using metadata statistics (occurrence of keywords):** The requestor is presented with a list of those GEMET concepts that have so far been used to annotate data (cloud representation). All other concepts are omitted. This ensures that each presented keyword results in at least one result when used in a search. The requestor also gets an impression of the influence a certain keyword has on the search results *even before he performs the search*: the larger a keyword, the more hits it will achieve. The main disadvantage of this approach is that it is based on exact matches of terms and does not include mechanisms for narrowing down or expanding the search results.

• **Using metadata statistics (co-occurrence of keywords):** This approach can be used to suggest additional search terms to the requestor to *narrow down* his search. After selecting one keyword, the requestor is presented with a list of keywords that are frequently used *in combination* with that keyword in the metadata. Again these keywords could be represented as a cloud according to their frequency in the metadata (figure 4).

Fig. 4: Suggesting additional search terms based on statistics for co-occurrence of keywords in metadata.

• **Using the GEMET structure:** This approach can be used to find additional matches by *expanding* the query with similar terms. The requestor is presented with a list of GEMET terms in a tree structure, possibly narrowed down by the selection of an INSPIRE theme. After the requestor has selected a search term, his query is automatically expanded to include similar terms. We currently investigate how a number of existing similarity measures can be applied to the GEMET structure. As first steps towards implementing these ideas we will focus on evaluating several similarity measures concerning their feasibility in combination with GEMET, and on collecting metadata statistics to be used for the mentioned approaches.