

# An integrated dataset of energy efficiency measures published as linked open data

Eva Blomqvist · Patrik Thollander

Received: 28 February 2014 / Accepted: 23 February 2015 / Published online: 20 March 2015  
© Springer Science+Business Media Dordrecht 2015

**Abstract** Despite an extensive energy efficiency potential, measures are sometimes not adopted due to barriers, such as lack of information. An integrated database of available energy efficiency measures, which has not existed previously, is one step towards overcoming such barriers. To address this, we present a dataset (i.e., data-base) integrating energy efficiency data from Sweden (from the Swedish Energy Agency) and the USA (from the Department of Energy's Industrial Assessment Centers), and publishing the data on the Web, using standardized Web languages and following the principles and best practices of so-called linked data. Additionally, several demonstration interfaces to access the data are provided, in order to show the potential of the result. These are entirely novel results, since this is the first dataset we are aware of that publishes this type of data using linked data principles and standards, thus integrating data from entirely different sources making them jointly searchable and reusable. Our results show that such data integration is possible, and that the integrated

dataset has several benefits for different categories of users, e.g., supporting industry and energy efficiency auditors in overcoming the information barrier for investment in energy efficiency measures, and supporting application developers to more easily integrate such data into support tools for energy efficiency assessment.

**Keywords** Energy efficiency audit · Energy efficiency improvement · Energy efficiency data · Linked data

## Introduction

Improving energy efficiency in industry is one of the most important means of reducing the threat of increased global warming. Despite an extensive energy efficiency potential, measures are sometimes not adopted due to various barriers, e.g., information imperfections and asymmetries. Overcoming the barriers needs new approaches to be deployed, one such being a database of available energy efficiency measures.

Increased global anthropogenic negative impact on our planet may be displayed by population growth, and number and impact of the used technologies (Ehrlich and Holdren 1971). The three major energy-using sectors are the residential and service sector, the industrial sector, and transportation. In the world, the residential and service sectors are the largest

---

This research was conducted within the DEFRAM project funded by the Swedish Energy Agency; for more information, see the DEFRAM website: <http://www.ida.liu.se/~evabl45/defram.en.shtml>.

---

E. Blomqvist (✉) · P. Thollander  
Linköping University 581 83 Linköping, Sweden  
e-mail: [eva.blomqvist@liu.se](mailto:eva.blomqvist@liu.se)

P. Thollander  
e-mail: [patrik.thollander@liu.se](mailto:patrik.thollander@liu.se)

energy user, while in developed countries like Sweden, the industrial sector is also in parity with the residential and service sector accounting for 35–40 % of primary energy use, depending on the economic situation (Swedish Energy Agency 2010). The industrial sector is thus of great importance to study in terms of improved energy efficiency, not only in Sweden but also world-wide. Improved industrial energy efficiency is according to IPCC, one of the world's most important means of reducing the global warming (IPCC 2007).

The energy use in the industrial sector is in general complex Waide and Brunner (2011). The single energy user, e.g., an electric motor, is in turn affected by the transmission, and the motor's energy use with transmission in turn affected by the whole electric motor system, be it pumping, compressed air, ventilation, etc. Waide and Brunner (2011). The single electric motor system is in turn affected by other processes in the company, e.g., the production processes. Summing up, industrial energy systems are to a large degree of great technical complexity and investments are often heavily capital intensive and thus more difficult to perform without thorough evaluation of the investment. Moreover, investments related to production processes often face higher degrees of risk, and thus need a thorough evaluation of the investment. Sharing already undertaken investments, their savings and investment costs may thus greatly help improve energy efficiency in industry through letting others validate their own planned investments, and inspire completely new investments and improvements.

## Motivation

Barriers to improved energy efficiency are often sprung from information-related barriers such as information imperfections and asymmetries, e.g., imperfect information, split-incentives, principal-agent relationship, etc. Sorrell et al. (2004) and Enrico and Trianni (2014). Also, large parts of so called hidden costs, e.g., search cost for the investment, are often mentioned as severe barriers to improved energy efficiency (Ostertag 1999; Bleyl and Eikmeier 2009). Other cited barriers are risks, lack of access to capital, and bounded rationality (Sorrell et al. 2004; Thollander and Palm 2013). Overcoming these barriers are of utmost importance in

improving energy efficiency in industrial energy systems. In summary, improving energy efficiency in industrial energy systems are of great importance in reducing negative environmental impact, but is challenging due to greater technical complexity than energy systems in the residential and service sectors (Thollander and Palm 2013). Moreover, it has for more strategic measures often extensive investment costs, many times associated with large risks, such as the risk for production disruptions etc. (Sorrell et al. 2004; Anderson and Newell 2004; Thollander and Ottosson 2008; Thollander et al. 2010). Furthermore, these strategic technology investments often face information-related barriers, and moreover extensive hidden costs or transaction costs (Ostertag 1999; Bleyl and Eikmeier 2009).

These challenges together call for new methods and approaches to promote energy efficiency in industry. One means that has emerged in the past three decades is various forms of online tools on the Web, e.g., the MURE database (Odyssey-Mure 2014), the US Department of Energy's Industrial Assessment Centers' (hereafter denoted IAC) database (IAC 2014), and various forms of energy audit protocols (e.g., Spice 3 (2014)), and energy management tools (Swerea Swecast 2014). One such online tool, a database<sup>1</sup> of previous improvements and their results, has been proposed as a tailored tool to help overcome information-related barriers, and reduce transaction costs and risks, leading to improved energy efficiency in industry (Cagno et al. 2010).

The idea of linked data (LD) has opened up a brand new way of structuring data and creating extremely advanced databases that are inherently Web-enabled through their graph-based link structure and the use of globally unique identifiers (URI:s) for data elements<sup>2</sup> (Berners-Lee 2009). In particular, the concept of *linked open data* (LOD), i.e., data that is published

<sup>1</sup>Throughout this paper, we will use the term "dataset" to refer to such a database. The reason for avoiding the term "database" is that it has historically been used to refer primarily to a specific database technology, i.e., relational databases, while in this paper, we deal with another kind of database, i.e., a graph database, which is commonly denoted a "dataset" when published on the Web.

<sup>2</sup>The technical principles behind linked data were presented by Berners-Lee (2009). For an updated view of the current linked open data cloud, see illustrations at <http://lod-cloud.net/> or the list of datasets tagged with LOD available from <http://datahub.io/dataset?tags=lod>

on the Web according to linked data principles, but which is additionally *open*, i.e., freely available to the public without charge, has become an enabler for a new “Web of data” (Bizer 2009). This means that the Web is more and more being used as a form of global data store, or database if you want, consisting of machine readable data instead of only web pages for human consumption. LOD thereby enables developers all over the world to develop software applications based on open data worldwide, without having to care about proprietary formats—instead, data is available through the standardised representation languages for linked data (Bizer 2009).

The aim of this paper is to present the motivation, structure, and logics behind a linked open dataset of real industrial energy efficiency improvement measures built on the basis of linked data principles. The solution presented in the paper is to the authors’ awareness unique, both since no dataset (i.e., database) of real industrial improvement measures has so far been published using linked data principles, but also since no database has previously attempted to integrate data from several countries, and even two continents. The results of the paper may be used to facilitate the construction of future industrial energy-related datasets, hence, our focus has also been on recording lessons learned and ideas on future work.

#### Novelty and benefits of the result

The published dataset will be useful to mainly three categories of users:

- Industries and energy auditing specialists
- Researchers and decision-makers
- Application developers

As mentioned above, the primary purpose of the dataset is to support industries, and auditing professionals working with those industries, to overcome the information-related barriers to energy efficiency that currently exist. For these users, being able to find and search the data in a simple but integrated fashion is an essential requirement. So far, most datasets that exist are (i) isolated “islands” of data, e.g., originating from one specific project, making them restricted in both time, focus area, and geographical coverage, (ii) presented in their own specific format, e.g., using proprietary formats such as Excel or PDF, and

implicit data models that are not semantically well-specified, which makes them hard to both search and integrate. We remedy this by showing how linked data principles can be used for data integration, and subsequently meet the requirements of this user group by providing a dataset that is freely available on the Web, including a demonstration search interface. This shows how easily one can build both general purpose and tailored search interfaces over the data, as soon as data is available on the Web in a standardized format.

However, not only the industries themselves may be interested in the data, but having easy access to an integrated dataset may also give new research opportunities, and new opportunities for decision-makers and policy-makers in various organizations to access and analyse that data more easily, to make better informed decisions. For instance, integrating (linking) data from Sweden and the USA, and making it available through a joint search interface, allows decision-makers to directly compare Swedish and US findings and recommendations, without first spending a lot of effort on “manually” integrating this data, e.g., in an Excel sheet. Finally, by providing the data through a Web service, we are enabling the direct use of data in other Web applications. Hence, application developers can call our service to retrieve data, taking advantage of our effort, for easy access to the integrated dataset. Potential applications envisioned to emerge could, for instance, be new support tools for energy auditing that rely on the data for recommending measures to be taken.

#### Energy-related datasets

To the best of our knowledge, the dataset that has been created in our project is unique in several ways, e.g., both in the technology used and in the way that it integrates and combines several different data sources from two different countries. However, there do exist related efforts that collect data about energy-related topics, and more specifically about energy efficiency, and make data available on the Web. In this section, we first present an overview of existing linked open data related to the energy domain, and then we discuss data particularly targeted on energy efficiency, including the data sources that were used to produce our results.

## Energy-related linked data

Numerous large datasets in various fields (e.g., biomedicine, cultural heritage, etc.) have been made freely available using the same linked data technologies as we propose in this paper, i.e., as LOD (Bizer 2009). Currently, we are aware of the following main efforts that have produced datasets that are part of the LOD cloud, which are concerned with the energy domain, although none are directly focused on energy efficiency audits:

- Enipedia
- Reegle
- Open EI
- EUROSTAT
- The World Bank
- The EnAKTing project

Enipedia<sup>3</sup> is a wiki website that collects data about energy-related issues, and in particular helps users to integrate, visualise, and analyze data (Davis 2012). Through the wiki templates, data is collected and expressed in linked data formats, and then displayed in different visualizations. For instance, the site contains information about energy infrastructures, such as natural gas supplies and global visualizations of electricity production using interactive maps.

Reegle<sup>4</sup> is a portal for publishing information about clean energy and energy efficiency projects. Data is displayed on a website for human browsing and consumption, but is also available through an API, structured according to the reegle ontology (i.e., a linked data vocabulary). The dataset contains, for instance, information about the energy consumption of various countries and their energy policies, as well as about energy efficiency improvement projects and their outcomes.

Open EI<sup>5</sup> is a US portal for collecting and displaying data about energy-related issues. Data mainly focuses on energy production, on a global and national scale, but also includes some data about energy consumption, by linking to survey data about smart meters and smart grid, energy consumption in buildings, as well as linking to the IAC database (IAC 2014).

A more general effort is publishing data from EUROSTAT as linked data<sup>6</sup>, which also contains energy-related statistics on a high level of aggregation (e.g., national energy consumption). Similarly, the World Bank<sup>7</sup> has started to publish some datasets, including data on climate change and energy.

Finally, two slightly older sets of data (from 2007) originates from a research project called EnAKTing, which was a pioneering project for producing LOD in the UK. The data concerns mostly road traffic energy consumption in the UK until 2007<sup>8</sup>, and greenhouse gas emission in UK regions until 2007.<sup>9</sup>

In summary, there are several general efforts for collecting linked datasets and other sources, e.g., making them available through various online portals on a global, European, and national level. However, most of the data is highly aggregated, e.g., on regional or national levels, rather than dealing with single organizations and energy efficiency improvement in those. To the best of our knowledge, there is no linked dataset present that publishes individual energy efficiency measures, implemented by companies and other organizations, their estimated costs and resulting energy savings, as LOD. The closest approach to ours is the Reegle portal, which, however, contains more aggregated data, focusing on energy-related *projects* and policies on a general level, without going into detail on their actual results in terms of measures coupled with their results in terms of saved energy.

It will be of essence to, in the long term, advertise our data in some of the abovementioned portals, but at the moment, the only related dataset that we were able to reuse and link to our Swedish data is the one from IAC, which is, however, not yet available as LOD. In order to be compatible with other datasets, we have nevertheless also looked into reusing the existing vocabularies (ontologies) that are used to express some of the data in the portals and projects mentioned above. We have therefore aligned our data model to Reegle, in order to make our data compatible with what exists in that portal. We will discuss this further in the “Publishing linked data” section of the paper.

<sup>3</sup><http://enipedia.tudelft.nl/>

<sup>4</sup><http://data.reegle.info/>

<sup>5</sup><http://en.openei.org/>

<sup>6</sup><http://eurostat.linked-statistics.org/>

<sup>7</sup><http://worldbank.270a.info/html>

<sup>8</sup><http://energy.psi.enaktng.org/>

<sup>9</sup><http://co2emission.psi.enaktng.org/>

## Energy efficiency data

As was noted in the last section, there is currently no linked open dataset that we are aware of, which covers energy efficiency measures and their energy savings. However, if one looks a bit broader, to other kinds of data collections, i.e., not only LOD, such data is available on the Web, but in various proprietary and non-machine-processable formats.

The IAC has the world's largest industrial energy-related database consisting of more than 116,000 measures (as of 2012 when this project was initiated) from real cases (IAC 2014). The IAC database is currently available through an online search interface, as well as downloadable Excel-files, i.e., it is primarily intended for human consumption through a graphical Web interface, and only available for download and reuse in the proprietary MS Excel format.

The idea of a large database has also been presented in a Swedish report by Söderström et al. (1994). However, to the authors awareness, no industrial energy efficiency-related database concerned with the end-use of energy still exist in the Nordic countries. Nevertheless, the Swedish Energy Agency is producing lots of data concerned with energy efficiency, but currently it is not published in a joint database or a linked dataset, but rather as a collection of documents downloadable from the website of the agency (in proprietary formats, such as PDF and MS Excel). Although this is in fact open data, it is not as easy to integrate, analyze, and reuse as it could be, without manually processing it in some way, due to the way it is made available.

Table 1 shows an overview of the datasets used as source data for our effort. PFE is a national energy auditing program in Sweden, and this dataset is the result of the first project period during 2004 and 2005. EKC is a financial incentive provided by the Swedish Energy Agency, where organizations may apply for co-financing of their energy audits, but in return have to report their findings and measures to the agency. For more explicit information about the PFE, please see Stenqvist and Nilsson (2012). For more information about the EKC, please see Thollander and Dotzauer (2010) and Karlsson and Thollander (2012). The IAC database is described in detail by IAC on their website (IAC 2014).

An overview of the content of the IAC, PFE, and EKC data sources is given in Table 2. The names

of organizations are codified in the IAC database, and for location we settle for the country in the US case. The assessed Swedish industries are classified into the Swedish industry classification system called SNI, which is the national version of the NACE Rev. 2 standard prescribed by the European Union. The IAC data instead uses the SIC system for industry classification. PFE only contains information of which year the assessment program was started, EKC only contains this information implicitly through the year when the co-financing was granted and the report was supplied, while the IAC recorded the actual assessment year. The type of measure is in the IAC dataset classified in a US-specific taxonomy called the Assessment Recommendation Code (ARC) (Muller and Kasten 2007), while the Swedish PFE uses its own (non-hierarchical) list of categories. EKC assessments are not pre-categorised at all. However, both PFE and EKC contain free text descriptions of the measures, while IAC settles for the textual titles of the respective ARC as a description. Pay-off time in PFE is calculated as straight payback without interest rate<sup>10</sup>, which is the same in the IAC data. Implementation details are in the Swedish datasets given through the year of actual or planned implementation, while the IAC instead have a categorisation of implementation status (into the categories of *implemented*, *pending*, *not implemented*, or *data not available*). Only the PFE data contains information on the method used for verifying the energy savings. In addition to what is listed in the table, each dataset also contains some project-specific data, such as who performed the assessment, etc., but this data was not considered relevant for our effort and is not discussed further in the paper.

In summary, we note that there exist several national data sources from different countries (e.g., Sweden and the US), which are freely available as open data. Additionally, we notice that datasets from Sweden and the US respectively have considerable overlap (see Table 2), although there are some differences in the use of taxonomies, units of measure, and formatting. However, none of these datasets have so far been published using linked data principles, hence, they are difficult to reuse and combine, due to that they mainly use proprietary file formats (e.g., PDF and MS

<sup>10</sup>In our data sources, pay-off time and payback time are both used, hence in this paper, we use them as synonyms.

**Table 1** Data sources used in our project (when downloaded in late 2012)

	PFE	EKC	IAC
Number of audited organizations	93	225	15 570
Number of measures in the dataset	1 256	1 438	116 960
Number of planned or implemented measures	1 216	968	54 474
Data collected during time period	2004–2005	2011–2012	1981–2012
Measures implemented during	2004–2011	2011–2014	(Unknown)

Excel) and target human consumption through graphical user interfaces. For making data truly valuable, and reusable, this data should be available in standardised open formats, and published on the Web in a manner that makes them not only consumable by humans but also directly accessible by software systems, e.g., external Web applications built by a third party. To the best of our knowledge, our dataset is the first to address this challenge, for detailed energy efficiency data. In the next section, we describe in detail the technologies used, and how our dataset was produced, so that others may follow a similar methodology and publish and link more data from other countries, in the end creating an energy efficiency LOD sub-cloud.

## Methodology

In this section, we first briefly introduce the technologies behind linked data (LD), in order to further clarify why this method for data publication was chosen, and what are its advantages. Next, we describe the methodology we have applied in the publication process, and then we describe some additional efforts that have given the dataset added value, including manually classifying the measures according to a uniform taxonomy, and linking the data to other datasets.

### Linked data—what is it?

The idea behind LD is that data is published in a standardized format, i.e., the resource description framework (RDF) (The RDF Core Working Group 2004), that provides a (globally) unique identifier for each data item, i.e., a so-called URI which is unique over the whole Web. This allows individual data elements to be uniquely identified, even over the entire

Web, and referenced, to express relationships between data, thus creating links between data. The links can relate data within a dataset, just as we can link elements in a traditional relational database, but now we can also link between multiple data sets, since the identifiers are (globally) unique over the entire Web. This allows us to place individual data, or datasets, in a context—we can link our data to other datasets that have overlapping information, or have already described the concepts that we are reusing, i.e., we can reuse other people’s data in a straight forward way. Linked data that is published openly on the Web, i.e., under a license that makes it free to reuse, is usually called linked open data, i.e., LOD.<sup>11</sup>

The principles of linked data were first summarized by the founder of the Semantic Web, (Berners-Lee 2009). He summarizes the principles in four bullet points that we quote here (Berners-Lee 2009):

1. “Use URIs as names for thing”
2. “Use HTTP URIs so that people can look up those names.”
3. “When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)”
4. “Include links to other URIs so that they can discover more things.”

To illustrate the linking property of LD, two examples that come from our published dataset are (i) information about places, such as cities and counties, and (ii) the industry classification codes used in Sweden (called SNI-2007, which is the national

<sup>11</sup>LOD applies the principles of LD, while additionally making the data open, i.e., freely available on the Web. In this section, we discuss the underlying principles of LD, which apply even if the data is not openly available.

**Table 2** Details of the content of the IAC, PFE, and EKC datasets (X indicating that the information is present in the dataset, and (X) that it can be derived or is partly available)

Content	IAC	PFE	EKC
Name of assessed organization	(X)	X	X
Location of assessed facility	(X)	–	X
Size of assessed facility	X	–	X
Industry classification of organization (SNI-2002)	–	X	–
Industry classification of organization (SNI-2007)	–	–	X
Industry classification of organization (SIC)	X	–	–
Program start year	–	X	–
Assessment year	X	–	(X)
Type of measure/recommendation	X	X	–
Free text description of measure/recommendation	(X)	X	X
Reduction of electricity usage	X	X	X
Reduction of other energy sources	X	–	X
Investment cost	X	–	X
Pay-off/pay back time	X	X	–
Planned implementation year	–	–	X
Actual implementation year	–	X	–
Implementation status	X	–	–
Method of verification of savings	–	X	–

Note that the IAC data includes a lot more information than what is included here, but in this project, we have focused on information that overlaps with information recorded in the Swedish datasets, or at least which is somehow transferrable or related to such information.

implementation of the EU standard NACE Rev. 2). On the Web, there is already data (in the standard RDF format) about Sweden’s municipalities and counties, as well as about the industry classification codes in the SNI-2007 standard. Both datasets are published by Statistics Sweden (SCB), as a result of a previous research project around linked data. By referring to URIs in these datasets, instead of creating our own identifiers for counties, municipalities, and SNI-codes, a user of the published data can, if desired, directly access additional information from Statistics Sweden about these data elements, such as the actual location of a municipality, or the description of the SNI-code. For example, suppose we publish data on energy savings in a company in the municipality of Uppsala. In our data, there is no information about the location of Uppsala, how many people live there, or the climate of the region. However, if this information is needed, e.g., for research purposes, or for building an application on top of our data, the links allow to directly retrieve such information from Statistics Sweden over the HTTP-protocol, i.e., without having to

first download and reformat the information to suit our needs.

A prerequisite for the above scenario to work on a Web scale, however, is that those who publish data use a standardized language<sup>12</sup> to express these data, and their global identifiers. Luckily, this standardization is already in place, through the W3C standards for using URIs to identify resources on the Web, RDF (The RDF Core Working Group 2004) to describe how these resources relate to each other, and the SPARQL query language (The W3C SPARQL Working Group 2013) to retrieve information from the data sets.

Finally, to make data more interoperable, it is additionally a best practice to publish your data model, along with the data itself. This allows anyone who

<sup>12</sup>Note that this applies to the technology to publish data—we still cannot necessarily guarantee that we “speak the same language” when it comes to what we mean by different concepts and data elements. The latter is of course also important, but is solve at a later stage, by publishing the data model, i.e., ontology, also using a standardized language.

finds the dataset to also explore its model, since it is explicitly defined and stored on the Web along with the data. Such data models are usually called “vocabularies” when used as simple data models for LD, but another common name (which sometimes indicates a more elaborate logical structure of the model) is “ontology.” To allow for the data models to be machine interpretable, e.g., usable by Web applications, they should be expressed using another standardized language, namely the Web Ontology Language (OWL) (The OWL Working Group 2009). For instance, in our case, we are publishing data about energy efficiency measures, their effects, and costs. If we simply publish the RDF data, without specifying the vocabulary, we will have a bunch of URIs (the identifiers of measures) and numbers (effects and costs) linked to them; however, there will be nothing to tell us what these numbers signify or what the links mean. It would be like publishing a table of data without column headings, and without a description of what the table signifies. Hence, the data model, i.e., vocabulary/ontology, tells us what each data item signifies, and what each relation linking the data elements mean. Furthermore, when publishing LD, such a vocabulary is not simply a text document containing a human readable description, or a term in a column heading, instead the model is described using a formal language (i.e., OWL) so that also systems and Web applications can interpret it directly.

### Publishing linked data

In order to publish our data as LOD on the Web, we had to go through three main steps:

1. Finding, or creating, the vocabularies (ontologies) to be used as the data model for the new LOD
2. Harmonizing and cleaning the source datasets
3. Transforming the source data into RDF and uploading it to a triple store

The first step can be seen as the process of creating the data model for the LOD. However, in contrast to when creating a data model for, for instance, a traditional relational database, when developing LOD, the best practice is to reuse as much as possible of already existing data models (existing vocabularies/ontologies) available on the Web. A common practice is to reuse parts (i.e., even single concepts

and relations) from existing authoritative vocabularies, even if the rest of that vocabulary may not fit our purpose. Hence, we started by listing the potential vocabularies for reuse, and picked relevant concepts and relations from each one that we intended to reuse. This process resulted in the decision to reuse concepts and relations from all the vocabularies in Table 3.

However, it turned out that after reviewing all available vocabularies (e.g., from the efforts mentioned in the “Energy-related linked data” section), there were still parts of the data, in particular the parts specific to the energy efficiency domain, that we were not able to represent. Based on this observation, we additionally created our own vocabulary<sup>13</sup>, specifically targeted at expressing information about energy efficiency measures, their costs, and effects. For instance, we had to add concepts representing the investment cost of a measure, and the effect (in terms of energy saved) of a measure, since these aspects were not found in existing vocabularies. An illustration of the vocabulary, including the reused concepts and relations from existing vocabularies, can be seen in Fig. 1.

Although the data from the original sources could be directly transferred to RDF at this point, without additional manual effort, we chose to partly harmonize and clean (by identifying and correcting errors) the data prior to publication, in order to make the dataset easier to reuse and to access in the future. With harmonization we mean, for instance, to as far as possible use the same metric prefixes for our basic units of measure, instead of mixing them as in the original data, e.g., to always use kWh instead of sometimes using kWh and sometimes MWh. Another harmonization is to always use the same format to represent certain types of literals, e.g., a date, in the data, etc.

Additionally, as can be seen in Table 2, even the two Swedish datasets are not using the same industry classification scheme (SNI-2002 is an older standard, while SNI-2007 is the current one), and the IAC data additionally uses the SIC standard. While translating from SNI-2002 to 2007 is quite straightforward, SIC and SNI are not equally compatible, i.e., lacking a one-to-one mapping even between some of their top categories. Producing a full translation of SIC into SNI

<sup>13</sup>The current version of the vocabulary is available here (in OWL-format): <http://www.ida.liu.se/projects/semtech/schemas/energy/2013/09/efficiency.owl>



**Table 3** Vocabularies from which we have reused concepts or relations

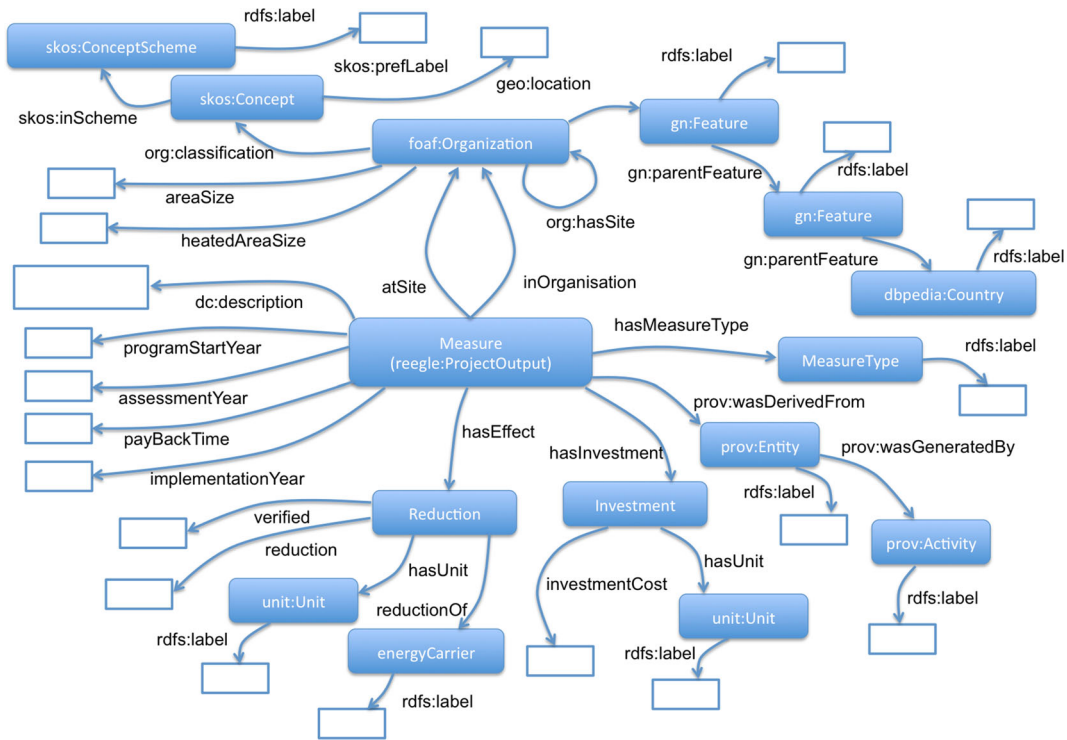
Name (prefix)	URI	Reused elements
Dublin Core (dc:)	<a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a>	dc:description
FOAF (foaf:)	<a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a>	foaf:Organization
Geo (geo:)	<a href="http://www.w3.org/2003/01/geo/wgs84_pos#">http://www.w3.org/2003/01/geo/wgs84_pos#</a>	geo:location
Geonames (gn:)	<a href="http://www.geonames.org/ontology#">http://www.geonames.org/ontology#</a>	gn:Feature gn:parentFeature
Org (org:)	<a href="http://www.w3.org/ns/org#">http://www.w3.org/ns/org#</a>	org:classification org:hasSite
Provenance (prov:)	<a href="http://www.w3.org/ns/prov#">http://www.w3.org/ns/prov#</a>	prov:wasDerivedFrom prov:Entity prov:wasGeneratedBy prov:Activity
RDFS (rdfs:)	<a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a>	rdfs:label
Reegle (reegle:)	<a href="http://reegle.info/schema#">http://reegle.info/schema#</a>	reegle:ProjectOutput
SKOS (skos:)	<a href="http://www.w3.org/2004/02/skos/core#">http://www.w3.org/2004/02/skos/core#</a>	skos:Concept skos:inScheme skos:ConceptScheme skos:prefLabel
Unit (unit:)	<a href="http://www.w3.org/2007/ont/unit#">http://www.w3.org/2007/ont/unit#</a>	unit:Unit

would be a project in itself, i.e., one could go down to more detailed levels of classification (rather than the top categories) and try to find mappings between these. Potentially, the problem may arise that there is no one-to-one mapping for some categories even at the most detailed level. In that case, it would be impossible to create an automatic translation script between the two classification systems, rather one would either have to re-classify individual organizations manually, using the other classification scheme, or settle for something less crisp than a strict one-to-one mapping. For instance, the linked data vocabulary simple knowledge organization system (SKOS) (W3C Semantic Web Deployment Working Group 2012) is a vocabulary expressing relations between knowledge organization systems, such as taxonomies, hence relations such as `skos:relatedMatch` could be used to instead link those categories in the respective taxonomies that are somewhat related, although not completely overlapping.<sup>14</sup> However, this would still take a substantial bit of effort. Considering these difficulties, we decided not to include such an in-depth study

<sup>14</sup>We are already using relations from the SKOS vocabulary for other purposes in this project (see Table 3), hence, this would be a natural step to take in future work.

of the industry classification systems in our project. However, one quite common use case (also a use case requested by our funding body, the Swedish Energy Agency) was to be able to search for “measures implemented in similar organizations to the one at hand,” which requires some way of specifying the type of organization. Therefore, despite the presented problems, we tried to find a way to at least accommodate this use case for parts of the data, if not all of it. Hence, we resorted to only providing SNI classifications for the top SIC categories that do have an obvious one-to-one mapping with SNI, and leaving the rest without a mapping, i.e., simply leaving those organizations we cannot automatically and accurately re-classify without any SNI classification, in order not to introduce any errors.

It should be noted that this harmonization is not a necessary prerequisite to transforming the data, we also carefully designed the vocabulary (in step 1) so that we preserve information such as units of measure, etc. This means that most of the data would be semantically well-specified even without this harmonization, however, for human consumption, e.g., entering search queries and showing data in search results, it is more intuitive to view data that is somewhat harmonized. Additionally, for the IAC data, we



**Fig. 1** An illustration of the overall vocabulary (ontology, or data model) used for our dataset. *Rounded boxes* represent concepts in our vocabulary, and *arrows* represent relations to other concepts or literals. Relations to literal values, i.e., numbers, text, etc., are represented as *arrows ending in an*

*empty box*. Namespace prefixes (i.e., what is written before the colon) indicate the provenance of the concept or relation, i.e., what existing vocabulary it comes from. Concepts and relations without a prefix are locally defined in our own vocabulary

have resolved a number of the codes that were used in the dataset, in order to get a representation that more closely resembles the one of our Swedish datasets. For instance, IAC uses alphabetical codes to express the energy source, e.g., electricity or natural gas, and each such source then is implicitly associated with a unit of measure. We have replaced those codes with explicit references to concepts (with human readable labels), representing the types of energy sources, and additionally included an explicit mention of the unit of measure related to each data value. This in order to reduce the memory load on a casual data user, not having to remember the meaning of the codes and which unit of measure is associated to what energy source.

In addition to the harmonization, we also took the opportunity to clean our Swedish datasets by checking samples of the data in particular for systematic errors. In some cases, systematic errors were discovered, and in these cases, we had to decide either to

correct the mistake (if possible) or to leave the erroneous data out of our dataset. One such example was the exact dates of implementation in the Swedish datasets, which seemed to contain many unrealistic dates, e.g., months and days that do not exist. This was due to the numerous different date formats that had been erroneously imported during data collection. In this case, we understood the problem, however, we did not find a suitable way to repair it, without going through all the records one by one and checking them against their written reports. This, together with the fact that this information is not absolutely essential to the dataset, i.e., we could still keep the year of implementation while discarding the exact date, led to the decision to discard the date information altogether, while keeping just the year. In other cases, there seemed not to be a systematic error behind the erroneous data items found, but rather simply a mistake in a manual calculation, or similar. In such cases, we decided to include or discard the data based on

mainly how valuable the information was considered to be. Thus, highly valuable categories of information, such as saved energy, or investment cost, was included even though we suspected that in a few specific cases (i.e., some individual data entries in this category fo data) there may be errors. The data from IAC was used as it is, without any attempt to discover and correct errors.

Finally, the datasets were transformed into RDF by using the RDF plugin<sup>15</sup> of the Google/Open Refine data transformation tool. This involves first expressing a transformation scheme in terms of how the data is mapped to the vocabulary created in step 1, and then the tool creates the RDF file automatically using an export functionality. In order to make the data accessible on the Web, we installed a so-called triple store (c.f. a database management system, but specific for LD<sup>16</sup>) on a research server at Linköping University, and uploaded our data onto this server. The server provides a SPARQL endpoint, which is a Web service that can be called via the HTTP protocol, to retrieve data in response to SPARQL queries. This service is freely available on the Web, but is intended for application access (machine access) and is therefore not intuitive to human users. To let interested parties access and explore our data, we therefore also created two Web interfaces for human users, one for LD experts, where they can write their own SPARQL queries and visualize the results, and one for non-expert users, where one can search the dataset by expressing a query through menu choices and checkboxes, developed as a simple Javascript application.

### Categorization and linking

Apart from the general error correction and harmonization mentioned above, we have made two major additions to the data:

- Classifying the measures (recommendations) into a shared taxonomy of measure types
- Linking the data to external data sources to make it part of the LOD cloud

The first addition, i.e., using a shared taxonomy of measure types, is important from an information

retrieval perspective, i.e., for end-users when trying to find information within the dataset. It allows a user to find all the measures of a certain type, without resorting to full-text search inside textual descriptions formulated by individual energy auditors (who would probably use very different terminology, making it hard to find a complete set of relevant answers to a user's search query). The latter, i.e., linking, is mainly important for supporting system developers and researchers when reusing the dataset for their applications or for performing research, i.e., it allows to automatically retrieve related information, which forms the context of our dataset.

Introducing a shared taxonomy, however, is not straightforward. The IAC data uses an extensive classification of measures (the ARC), describing on a detailed level the type of measure that was recommended. ARC is a hierarchical classification scheme (taxonomy) of measures that ranges from general areas down to specific measures, such as replacing a specific part of a machine or facility (Muller and Kasten 2007). Although this extensive taxonomy would be possible to use also for the Swedish data sources, we chose not to do so for a number of reasons. First and foremost, structuring the data using 900 pre-defined measures after the measures have been reported, would be a very thorough job, taken the extensive amount of data. Moreover, the IAC data is related to small- and medium-sized enterprises (SMEs) while the Swedish PFE is related to majorly large energy-intensive companies. This motivated the choice of a more general taxonomy, i.e., the one developed by Söderström et al. (1994), as it is general and more applicable to datasets coming from various industrial sectors. In order to translate the ARC-categorization to our taxonomy, a complete mapping from ARC categories to the (Söderström et al. 1994) taxonomy was developed, and each IAC recommendation was thereby given an additional category also from this taxonomy. The ARC textual descriptions were kept as descriptions of each measure in the IAC dataset, since that data otherwise lacked free text descriptions of the recommendations. The EKC data lacked any kind of classification in its raw form, hence, these measures had to be classified manually. Finally, the PFE data used a related but slightly different classification, compared to the (Söderström et al. 1994) one, but here we could apply a semi-automatic translation since we were able to first find

<sup>15</sup>Tool can be downloaded here: <http://refine.deri.ie/>

<sup>16</sup>In our case we used the open source Sesame RDF store that can be found here: <http://www.openrdf.org/>

a one-to-one mapping between categories (although some problematic entries in the original classification had to be corrected manually). Without this joint taxonomy, the user would have had to be aware of ARC to find results from the IAC dataset, to be aware of the PFE taxonomy (specific only to that project) for finding results also from this dataset, and finally perform a full-text search using a set of terms in order to find measures based on their textual description within the EKC dataset. It should thereby be quite obvious to the reader how such a joint taxonomy can improve the findability of measures. It is now enough to choose one of the categories in the joint taxonomy as search query, and one will be certain to receive *all* the related measures, both from the IAC, PFE, and EKC datasets, related to that category of measures, e.g., related to production processes, administration, or facility heating, for instance.

The second addition to the data, the linking, is made possible by the way we are publishing the data, i.e., as LD. As mentioned in the “[Linked data—what is it?](#)” section, one of the main features of LD is the (globally) unique identifiers given to data elements, which in turn makes it possible to refer to data from anywhere on the Web. On the one hand, this facilitates others to reuse and refer to our data in an efficient manner, but also allows us to reuse, i.e., link to, existing data on the Web, rather than replicating that data in our own dataset or simply not caring about mentioning the relations. We have used such links in two cases in our dataset, i.e., when referring to:

- The SNI-2007 codes
- Geographical locations

The SNI standard for industry classification is provided by Statistics Sweden (SCB), who are also providing these codes as LOD on their website. This means that each SNI code has its own URI, under the SCB domain, where SCB is providing additional information about that particular code, such as its textual description, its parent codes, and sub-codes in the taxonomy. Due to the reuse of those URI:s in our dataset, anyone who needs more information about a SNI-code can directly follow its URI and retrieve accurate and up-to-date information about it directly from SCB, rather than some (potentially outdated) information replicated in our dataset. Similarly, it would be easy to, for instance, replace all the detailed classification codes with their top-level category, e.g.,

in a user interface with a drop-down list letting the user select an industry category to look at, even though that information is not actually present in the data, but it can be looked up with a simple request to the SCB online data site. In fact, we are using this in our own Web search interface, for letting the user first select the main industry category, before having to specify the detailed classification code, although the top-level classification codes are actually not present in the data.

Similarly, we are linking to several data sources concerning geographical locations. In addition to the SNI-codes, SCB also provides an authoritative dataset listing all the municipalities and regions of Sweden, with unique URI:s, and providing additional information and links for all of them, such as their names, location in Sweden, official website, etc. Whenever information about the location of an assessed facility is available in our data, we therefore reuse (link to) those URI:s to identify the location, e.g., the municipality and region. Additionally, for more general geographical concepts, such as countries, we link to one of the most important data sets in the LOD cloud, namely DBpedia. DBpedia is a cross-domain dataset, based on the Wikipedia website, which provides URI:s for, and information about, anything that is mentioned in (at least the English) Wikipedia. DBpedia is also a “hub” at the centre of the LOD cloud that links together most of the available datasets, which makes it an important dataset to relate to. Hence, by linking to the DBpedia URI of Sweden (and the US), one could retrieve any information needed about these countries that is present in DBpedia, or other datasets linked to DBpedia, based on resolving that URI.

Similarly to the inclusion of SCB URI:s for SNI codes, these links do not have an immediate value within the dataset, instead the value emerges when someone wants to reuse our dataset for research purposes, or for building a software application on top of it. For instance, consider the use case of researching energy efficiency measures implemented by industries residing in small towns, and comparing these to what is done in bigger cities. The size of the city where the industry resides is not part of our dataset, nor should it be—it is clearly out of scope for us. However, by including the link to another authoritative dataset about municipalities, this information can immediately be retrieved, just by following this link and retrieving the additional information that resides on the SCB servers. The resulting data would

effortlessly be a combination of part of our dataset, and parts of the SCB data on municipalities, including the size of their towns, i.e., this is what we mean by putting our data into context, whenever necessary. Another use case would be to plot the data in a map interface, for users to explore. In this case, location coordinates can be directly retrieved through the links to SCB.

It should be noted that this linking is not strictly necessary for publishing our data, but it is part of the best practices for publishing LD (as mentioned in the “[Linked data—what is it?](#)”) section, and additionally it is an easy step towards making our dataset more usable, and reusable by others, since the links puts our data into a context on the Web, and would allow others to use the links when creating applications on top of our data.

## Results

Being a novel project for publishing linked energy data in Sweden, this project has resulted both in the published dataset itself as well as a number of lessons learned, including known issues of the current dataset, that leads to a number of future work opportunities.

### Dataset and demonstration interfaces

The resulting data has been published on the Web, and is available through a Web service, i.e., a so-called SPARQL endpoint. It can be accessed directly through the endpoint (mainly for machine access) or through either of our two demonstration user interfaces. The first user interface is targeted towards LD experts<sup>17</sup>, and application developers, who can explore the complete dataset and its associated data model through formulating and issuing SPARQL queries against the data. The second demonstration interface<sup>18</sup> is a simple Javascript application, which provides a guided way to formulate queries against the data, using menus and checkboxes.

<sup>17</sup>Available here:

<https://www.ida.liu.se/projects/semtech/energy/snorql/> We have reused and modified the open source SNORQL interface available here: <https://github.com/kurtjx/SNORQL>

<sup>18</sup>English version available here: <http://www.ida.liu.se/projects/semtech/energy/demo/index.en.html>

An illustration of the latter interface can be seen in Fig. 2. The set of checkboxes at the top represent the different data elements, i.e., information about a measure (recommendation), that a user could be interested in viewing. By selecting a set of these checkboxes, the user expresses what columns should be included in the result table. Below the checkboxes, a set of menus are available that represent ways to restrict the query, e.g., by viewing data for a specific industry category (SNI-code), a specific type of measure, a specific geographical location, or restricting the query to give a maximum number of results. The button entitled “Submit search” initiates the search (alternatively, the user can show the SPARQL query generated by their selection through the button “Show SPARQL” and, for instance, copy it to the expert user interface where it can be modified and re-issued against the data). An example search result can be seen in Fig. 3.

It should be noted that this is a general demonstration interface, i.e., it is not targeted towards any specific task, or user group, but serves mainly the purpose of allowing us to demonstrate the project results to domain experts in the energy field. Finally, the data is available for download, as an RDF dump<sup>19</sup>, to enable developers to set up their own triple store, or import the data into their tool of choice.

Looking at the data in more detail, one can see how the data model is reflected in the LD. Figure 4 illustrates how one measure (recommendation) is represented within the dataset, using the general model that was illustrated previously in Fig. 1. Following the LD principles, each instance of a concept is identified by a URI, i.e., a globally unique identifier over the Web. In the figure, the URI:s referring to the local namespace of our dataset have been abbreviated (using “...”) for readability reasons, while the links to external datasets are shown with full URI:s. The figure also illustrates how using LD standards and principles renders a graph of links between data elements, rather than the usual relational data model (e.g. usually illustrated as a table). While this is a data model that suits the Web, when visualizing data, e.g., as a result of a query, one can of course visualize it in a table format (as shown in Fig. 3) if desired. In Table 4, we also summarize the data that is included

<sup>19</sup>Current release available here: [http://www.ida.liu.se/projects/semtech/energy/Energy\\_201309.zip](http://www.ida.liu.se/projects/semtech/energy/Energy_201309.zip)

Choose what information about a measure (recommendation) you want to display:

Company- and facility-code     
  Industry classification (SNI-codes)     
  Geographical information  
 Data provenance (e.g. PFE or IAC)     
  Measure (recommendation) category     
  Measure (recommendation) description  
 Reduction of energy use     
  Pay-off time and investment cost     
  Planned/implemented (date)

Restrict your search by choosing the categories you are interested in below:

Industry classification (SNI-code)

Concerned process (measure category)

Geographical limitation

Limit the number of results

**Fig. 2** An illustration of the simplified search interface built for the dataset

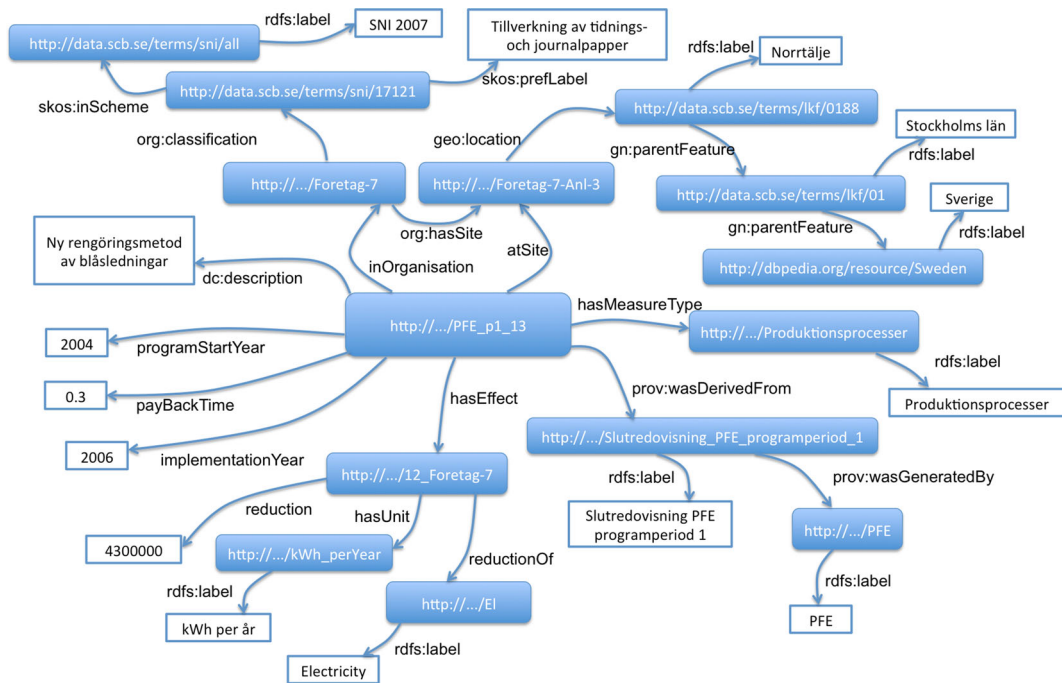
in the final (linked) dataset and briefly summarize how the source data from each data source was transformed (if needed) before included in our new dataset.

Some parts of the dataset have been omitted from the table due to space reasons, e.g., the provenance information stating from which original data source

SNI-code	SNI-description	Region	Measure category	Measure description	Energy carrier	Energy reduction	Unit	Provenance	Pay-off time	Investment cost	Unit
14	Tillverkning av kläder	USA	Tryckluft	2.4231 REDUCE THE PRESSURE OF COMPRESSED AIR TO THE MINIMUM REQUIRED	Electricity	15552	kWh	IAC	0.02	28	USD
14	Tillverkning av kläder	USA	Tryckluft	2.4231 REDUCE THE PRESSURE OF COMPRESSED AIR TO THE MINIMUM REQUIRED	Fuel oil	40	MMBtu	IAC	0.02	28	USD
14	Tillverkning av kläder	USA	Tryckluft	2.4221 INSTALL COMPRESSOR AIR INTAKES IN COOLEST LOCATIONS	Electricity	17921	kWh	IAC	0.26	460	USD
14	Tillverkning av kläder	USA	Tryckluft	2.4221 INSTALL COMPRESSOR AIR INTAKES IN COOLEST LOCATIONS	Fuel oil	46	MMBtu	IAC	0.26	460	USD
14	Tillverkning av kläder	USA	Produktionsprocesser	2.4111 UTILIZE ENERGY-EFFICIENT BELTS AND OTHER IMPROVED MECHANISMS	Electricity	19314	kWh	IAC	0	0	USD
14	Tillverkning av kläder	USA	Produktionsprocesser	2.4111 UTILIZE ENERGY-EFFICIENT BELTS AND OTHER IMPROVED MECHANISMS	Fuel oil	50	MMBtu	IAC	0	0	USD
14	Tillverkning av kläder	USA	Lokalvärme	2.7241 INSTALL OUTSIDE AIR DAMPER _ ECONOMIZER ON HVAC UNIT	Electricity	21096	kWh	IAC	1.68	3840	USD
14	Tillverkning av kläder	USA	Lokalvärme	2.7241 INSTALL OUTSIDE AIR DAMPER _ ECONOMIZER ON HVAC UNIT	Fuel oil	89	MMBtu	IAC	1.68	3840	USD
14	Tillverkning av kläder	USA	Produktionsprocesser	2.4151 DEVELOP A REPAIR_REPLACE POLICY	Electricity	27140	kWh	IAC	0.26	651	USD
14	Tillverkning av kläder	USA	Produktionsprocesser	2.4151 DEVELOP A REPAIR_REPLACE POLICY	Fuel oil	74	MMBtu	IAC	0.26	651	USD

**Fig. 3** An illustration of how results are presented in the simplified search interface, in this case data originating from the IAC database. The first two columns represent the industry classification of the assessed organization. The fourth column represents the classification of the recommendation in

terms of our Swedish taxonomy of measure types, while the fifth one shows the original description in terms of the ARC taxonomy. The columns further to the right then show the reduction of different energy sources, pay-off time, and investment cost



**Fig. 4** An example of how the data model is used to express the data, for one specific measure (*recommendation*) in the dataset. The example measure comes from one of the Swedish sources, hence labels are unfortunately in Swedish

the data came; in the table, we focus only on the most essential, and domain relevant, features in the data.

#### Lessons learned

A major focus area of this project was to explore the possibility of publishing energy efficiency assessments and recommended measures as LOD on the Web, and in particular to explore the combination of several datasets, both nationally and internationally. To this end, we have learned that this is actually possible, with a reasonable amount of manual effort. The manual effort was mainly needed in the steps for harmonizing and cleaning the data sources, as well as creating the mappings between standards (e.g., different industry classifications and different taxonomies of measures); however, we actually used slightly less effort for these tasks than expected, which is a positive experience. On the other hand, we have avoided the most difficult parts of the mappings, e.g., only creating a mapping between SIC and SNI where obvious one-to-one mappings exist, and leaving the rest of the categories unmapped, rather than going

into detail and judging the data case by case or modelling the relations through SKOS. Nevertheless, even with this conservative strategy, the mapping scheme covers about 73 % of the overall SIC scheme (on the two-digit level), meaning that 64 % of the IAC data (amounting to well over 75,000 recommendations) has received an additional SNI classification. Hence, if this is an acceptable level of coverage, our method can easily be applied again on new data automatically, without having to modify the mapping scheme.

Despite our efforts to harmonize and clean the Swedish datasets, we have noted that the data is still not completely correct in all cases. For instance, the following kinds of errors have been discovered:

- Wrong industry classifications
- Wrong or missing geographical location
- Unrealistic and missing data values (e.g., saved energy or investment cost)

Since the industry classifications is a commonly used search criteria, e.g., for an industry to find measures implemented by “similar” organizations, it is

**Table 4** Details of the origin and representation of the data included in the dataset

Content	Concepts and relations	IAC	PFE	EKC
Assessed organization	foaf:Organization in:Organization	URI (generated from the organization ID)	URI (generated from codification of organization name)	URI (generated from codification of organization name)
Location of assessed facility	geo:location gn:Feature	Fixed value (URI from DBpedia representing the US)	When specific location was present in the data: URI of the location from SCB, otherwise fixed value (URI from DBpedia representing Sweden)	When specific location was present in the data: URI of the location from SCB, otherwise fixed value (URI from DBpedia representing Sweden)
Size of assessed facility	:areaSize and :heatedAreaSize	Plant area as value of :areaSize	-	Values directly representing :areaSize and :heatedAreaSize
Industry classification of organization	org:classification and skos:Concept	SIC translated into SNI-2007 (if possible)	SNI-2002 translated into SNI-2007	SNI-2007 available
Program start year	:programStartYear	-	Year as present in data	-
Assessment year	:assessmentYear	Year as present in data	-	(Not included at present, since it has to be derived from the year of the economic support)
Type of measure /recommendation	:hasMeasureType and :MeasureType	ARC reclassified into joint taxonomy	PFE classification reclassified into joint taxonomy	Assessment manually classified into joint taxonomy
Free text description of measure	dc:description	ARC label	Textual description present	Textual description present
Reduction of electricity usage	:hasEffect and :Reduction, amount of reduction using :reduction, :hasUnit and unit:Unit, together with fixed value for :reductionOf and :EnergyCarrier	Energy savings where the code represents electricity	All energy savings	Energy savings representing electricity



**Table 4** (continued)

Reduction of other energy sources	:hasEffect and :Reduction, amount of reduction using :reduction, :hasUnit and unit:Unit, together with value for :reductionOf and :EnergyCarrier representing the specific energy carrier	:hasInvestment and :Investment, value stored through :investmentCost, :hasUnit and unit:Unit	:payBackTime	:implementationYear (currently not included)	:verified	Energy savings with other codes than electricity	Energy savings not representing electricity
Investment cost		Investment cost in the unit USD	Investment cost	Investment cost in the unit SEK			
Pay back time		Straight pay back time	Straight pay back time	Year reported			
Implementation year		-	-	Year reported			
Implementation status		(Implementation status)	(Implementation status)	-			
Method of verification of savings		-	-	Textual description of method			

Concepts and relations that are listed are the main concepts and relations needed to represent this particular part of the data. Additional properties, such as rdfs:label may also be used but are omitted due to space restrictions (the complete set of properties in our data model can be seen in Fig. 1). For each data source we give a brief description of what data was included, and if and how the data was transformed.

important that these classifications are correct and up-to-date, which does not seem to be the case for all reported data (in particular in the self-reported data of the EKC data source). A better method, rather than simply trusting the self-reported classification code, would be to directly retrieve this classification from the national Swedish register of organizations, maintained by the Swedish tax authority.

Geographical location is another important search criteria, but today, the locations are sometimes wrong or inconsistent. For instance, it may be that the organization has reported the location as the region of the facility that was assessed, but at the same time as the city where the head office is located, resulting in inconsistent data when the given city is not actually located in the same region as the assessed facility. However, with some additional effort, the locations of facilities could many times be verified against the national register of organizations.

Finally, in some cases, even important data values such as the amount of energy saved, or the investment cost, or pay-off time, may also be missing or contain errors. Some of these errors, in our Swedish data sources, have been introduced through human error when transferring the data from the original (textual) reports written by the assessed organizations to our data source (an Excel-sheet which was then transformed into RDF). Others are already present in the original reports. For these types of errors, one has to resort to manual effort for error checking, e.g., by reading the reports, recalculating values, and assessing their reliability. Recognizing the importance of data quality, addressing these three types of errors is the current topic of a follow-up project that was recently started and will be carried out throughout 2014. However, due to that this is a manual process, and errors are discovered and corrected manually, we are not able to alert users on potentially erroneous values before they have actually been manually checked (and once they are checked they are also manually corrected, or removed in case no correct solution can be found).

Notably, in this follow-up project, we have primarily found data quality problems in the dataset from the Swedish energy audit program (EKC). As a consequence, the research group now carefully goes through all energy audit reports of the Swedish energy audit program, i.e., the original source of data, in order to strengthen reliability of the data, and improve the

validity of the database. Without making an understatement, this is an extremely costly way of quality checking the data. If a company, government, researcher, or sector organization would initiate a data collection, with the primary aim of creating a harmonized database, this research paper provides some findings which are of general nature and can be used to improve such a data collection process. First, prior to the data collection, a harmonized taxonomy, expressing how to categorize data, should be decided upon (preferably an existing one should be reused, such as ARC or the one by Söderström et al. (1994)). In addition to this, harmonized data collection templates should be created, giving the auditors (data collectors) clear guidelines on what data that must be collected, and in what format it should be input. This template should preferably be accompanied by criteria for auditors, and an educational course on how the data should be collected, e.g., through metering, company sub-metering system, etc.. Finally, the created database should explicitly state from where data has been gained, e.g., through metering, sub-metering system, calculations, in order to display for the database user, the variety in reliability which the data displays.

While the problem of complementarity and comparability between different datasets from different energy programs is a national and international challenge facing not the least researchers in the field of industrial energy efficiency policy program evaluation (e.g., lack of harmonized calculation standards makes evaluators use different lifetime of the same equipment etc. leading to incomparable results when calculating net present value and cost-effectiveness), as regards complementarity of the collected data from the three different programs, a few words of caution must still be explicitly addressed. First, the American IAC and the Swedish EKC primarily targets small- and medium-sized companies, and the data originates from conducted energy audits. Previous research from Europe found that the American database from the IAC was transferable to European SME conditions (Cagno et al. 2010). With the developed transformation key, the two types of datasets may be seen to hold high complementarity. The third type of dataset, the Swedish PFE primarily address energy-intensive companies, mostly large, from which the top management often have dedicated staff solely

for energy management, energy monitoring systems in place, etc. The data from PFE originates from both the conducted energy audit within the program and activities emanating from the certified energy management system the companies need to install. Moreover, incorrectly reported data from the companies could result in them losing the tax exemption the program provides, and even juridical consequences. This makes the reliability of this data higher. However, thanks to the developed transformation key, the dataset may still have high complementarity to the other two datasets. In future national and international programs, it is strongly suggested to work on developing an international industrial energy efficiency data standard, where also a classification scheme for the collected data is created. Such standard and classifications would enable even higher comparability and various datasets could complement each other even further.

Multilinguality of the data is another issue that we have not completely resolved. Obviously, Swedish data will be expressed in Swedish and data from the IAC in English. At the moment, our graphical demonstration user interface does not distinguish between the two languages, hence search results may be mixing between the languages, in terms of descriptions and comments in either language being presented in the same result table (as can be seen in Fig. 3). In the dataset, we are able to express the information of what data is expressed in what language, through so-called language tags, and it is easily possible to include several translations of a term or a textual description. By applying LD principles, and providing URI:s for important concepts, we are thereby reducing the effects of multilinguality in data. For instance, we are able to determine that “electricity” (in English) is the same as “el” (in Swedish) since these are simply two different labels of the same individual in the data, identified by a URI rather than a term in either language. Still, for increasing the reusability of Swedish data in an international setting, such as the Web, one would like to also provide translations of the labels and descriptions of measures that occur in the data, which would again require some manual effort. Hence, for now, we have decided to show all the data, no matter what language it is in, through our user interface, but data translation is certainly an issue for future work.

An even more interesting issue when publishing and using data internationally, is its applicability to local conditions. As mentioned before, existing research results have shown that in general the database from the IAC was transferable to European SME conditions (Cagno et al. 2010). However, this may or may not be the case when the direction is the opposite, i.e., transferring data from Swedish conditions and using it to aid energy auditing in, for instance, the USA. Although, our experience in this field, and the relative compatibility of the data itself, as discussed above, leads us to believe that data would be applicable, it has not yet been shown empirically through some case study. Simple tailoring of data, such as transforming figures and presenting data in the unit of measure most prominently used in the country in question, could easily be handled. Either through storing such alternative values and their units of measures in the dataset, or handling the transformation online in an application interface. However, more subtle adaptations and tailoring may be harder to perform. For instance, it may be interesting to put investment costs and savings into context. At the moment, investment costs are presented in the local currency (with the currency clearly marked); however, that does not really tell a user from another part of the world much about how expensive it would be for his organization to do the same today, or what they could actually save, in terms of payback time and saved energy. Hence, in the long term, putting this dataset into context, when it comes to exchange rates and energy prices, of the countries where data originated, and at the time the assessment was performed, would also be a valuable means of internationalising the data.

#### Potential of the dataset

The potential and use of the created dataset differs from a conventional (relational) database in a number of ways. Firstly, it does not confine the data to reside within a specific company, trade organization, or national energy agency, but since the data resides on the open Web, this opens up for a brand new energy service segment, where, e.g., software for energy audits and energy management can utilize the dataset and create software applications directly linked to the database. Moreover, the creation of a

similar dataset for developing and small countries may be too optimistic. By sharing, openly, the dataset on the Web, measures are easily spread and data can be utilized even by auditors and authorities in countries that have no data collected on their own, and no resources to do so. Perhaps the most important advantage of the dataset, however, may be that it will help improve future energy audit programs by allowing for developing tailor-made protocols for auditors, and provide the possibility for energy program evaluators to compare their findings to other programs in other countries.

As mentioned at the beginning of this paper, the published dataset is targeted mainly at three categories of users, i.e., (i) industries and energy auditing specialists, (ii) researchers and decision-makers, and (iii) application developers. As a result of integrating three different datasets, which were previously only accessible separately, and in proprietary formats (such as PDF and MS Excel), we have shown how data integration can help increase the coverage of datasets, both in terms of time, geographical coverage, and types of industries that are in focus, as increasing the usefulness and reusability of the data. By integrating the IAC data with our Swedish data, one can now search for data as far back as the 1980s, and for industry categories where no Swedish data currently exists, since the IAC data is much more extensive than the current Swedish data. For instance, SNI 14 is the top category signifying industry that manufacture clothes, and for this complete category of industries, there is no data at all in the Swedish data sources, while there is plenty of assessments done on industries in this category in the IAC data, which can now be utilized as inspiration for similar companies in Sweden. On the other hand, we still find it valuable to include local data, which may be more tailored to Swedish conditions, and investment costs that are realistic under local conditions, etc.

However, in addition to simply increasing the coverage of data, over several dimensions, the way we have amended data, by adding mappings and new classifications, and the way we have published it also provide some benefits. It is now possible to retrieve (a large part of) the IAC data using the Swedish SNI-2007 classifications as a search query, and it is possible to query the complete dataset using one single taxonomy of measure categories, neither of which was possible before. Due to the publication as linked

data, it is also now possible to query all this data jointly, through one single query, and since we are querying the linked dataset, we can use all the parameters included to specify our query. Hence, we could, for instance, ask for the average energy saved through measures in production processes in the USA vs. in Sweden, just with one single query, and no preprocessing of the data. Of course, this would have been possible also before, but only by downloading all the IAC data to a local Excel sheet, selecting all the ARC-codes that represent production processes (based on a screening of the codes), calculating the average energy saved through those measures by writing a formula in the Excel sheet, and then doing the same thing (download, select the appropriate measures from a diverse set of categories, and making the calculation) for the Swedish PFE data, before finally comparing the resulting numbers. EKC data was not even present as a downloadable Excel-sheet prior to this project. Additionally, we could now even start exploiting the links in the linked data for asking queries. If we allow for federated querying, including also the datasets that we link to from our data, one could ask queries such as what measures are most frequent at facilities in large cities, compared to in rural areas? Or, what are the measures industries of my category, located in adjacent cities to mine, have performed, and what did those investments cost them?

Again, returning to the three groups of users, future work for further supporting the first user group would be to improve the online search interface, to target more specific users, e.g., separating the needs of industries looking for general inspiration from similar organizations, from the needs of highly skilled energy auditors, for instance. It should also be noted that the current user interfaces are only demonstrations, i.e., they are not optimized for a high search load and we cannot guarantee the uptime of the service. Transferring the service to a production environment, dimensioned for a realistic estimated load, would be necessary before such a service can be used as a reliable source of information for all of our Swedish industry.<sup>20</sup>

<sup>20</sup>We would like to point out that these performance issues are not due to the linked data technologies or systems used, but are simply an effect of this being a demonstration application where performance was not in focus, running on servers in the university's research environment, intended only for demonstration projects, not deployed services although open to the public.

For providing researchers and policy-makers with information for analysis and decision support, the service is most likely reliable and fast enough already; however, here the quality of the data may become a problem instead. When considering individual measures (recommendations), we expect industries and auditors to be able to use the data merely as examples and inspiration, rather than as “ground truth,” while researchers may rather use the data to create aggregations and statistics over data, making it harder to discover potential errors in individual data elements that may very well affect the overall results. This makes quality checking of the data a high priority, which led to our ongoing follow-up project dealing with this issue (mentioned in the previous section).

Finally, for the third user category, i.e., application developers, the technology by which we have published the data provides the greatest potential. Two main advantages can be observed by application developers, i.e., that the data is (i) represented using standardized formats, such as RDF and OWL, making it directly accessible and usable by development tools and libraries, without additional effort on the part of the developer, and (ii) available through a Web service, allowing for direct access and querying over the Web. The latter means that an application built on top of this data does not necessarily have to contain a local copy of the data, but can instead issue queries against the Web service, in order to retrieve the data needed by a certain user. This comes with several benefits, such as that the application does not have to maintain its own large data storage facility, it only has to manage the small amount of data relevant in each case, and the data will always be up-to-date, since it is retrieved “on-demand” from the service.<sup>21</sup> As mentioned already when discussing the first user group, this would, however, require a migration to a production server environment, rather than our current demonstration one.

<sup>21</sup>Note that this does not mean the dataset residing on our servers would only contain the latest additions to the data, it will always contain the complete set of data, incrementally built through additions of the latest data as it is loaded into the dataset. However, when a query is sent, the latest data, in addition to older data, is always available for retrieval, which may not be the case for a local copy that was downloaded some time ago.

The applications for the presented dataset was mentioned to be threefold. The Swedish part of the dataset has so far been used in a governmental ex-ante study on the effect of a future Swedish Technology procurement policy program for industry (Franck et al. 2014), in ongoing research, national and international (Thollander et al. 2014), e.g., in an IEA-project, ex-post evaluation of the Swedish energy audit program, and studies on district heating potential in Swedish industry (Djuric Ilic and Trygg 2014). It should be noted that the cited works used the Swedish data, thus primarily showing the usefulness of a dataset as such, and also the usefulness in having a uniform taxonomy, or standard, in categorizing data, that enabled program comparison (Franck et al. 2014). However, in, e.g., energy program evaluation, implemented measures could now, with our created dataset, be compared on a bottom-up level, something which up until now has not been possible in policy program evaluation.

## Conclusions and future work

In this paper, we have presented a dataset (i.e., database) integrating energy efficiency data from Sweden and the USA, and we have described how it has been published as linked open data on the Web, as well as being made available through several demonstration interfaces for accessing the data. This is the first dataset we are aware of that publishes this type of data, using linked data principles and standards. Our results show that such data integration is possible, and indicate that the integrated dataset will provide several benefits for different categories of users, e.g., supporting industry and energy efficiency auditors in overcoming the information barrier for investment in energy efficiency measures that has been shown to exist, and supporting application developers to more easily integrate such data into support tools for energy efficiency assessment.

One of the strongest contributions of this paper however, may not lie in the actual creation of the database but in the applications of such. For example, the centuries much debated size of the energy-efficiency gap may be partly resolved by displaying openly the results from conducted energy audit programs from various countries and regions. As for today, the magnitude of the energy

efficiency gap is not modeled bottom-up but top-down, mostly by energy efficiency experts. Moreover, as the used technology is open, the transferability to regions with lower amount of data, e.g., developing countries where the creation of a database may be more problematic, may lead to improved energy efficiency far beyond the actual energy audit program. Finally, the use and application of databases for the building sector and for policy measures (e.g., the MURE) is currently already in place; it is of utmost importance for the scientific community to support the creation of such databases also for the industrial sector of the economy. This paper presents one small step for research in this direction.

Future work opportunities, in addition to the ongoing efforts mentioned in the last section (e.g., improving the technical facilities for data access, and improving data quality) include to more comprehensively address the mapping problem, e.g., by creating a complete mapping between different industry classifications and to address the multilinguality problem. Furthermore, to introduce a more detailed provenance trace for our data would be a useful addition, for instance, to be able to explicitly record inside the dataset itself when a value was corrected (e.g., due to an error that was found), what was the previous value, etc. On a broader note, for the future, we envision a network of linked datasets, similar to our current result, where a multitude of countries contribute their own energy efficiency assessment data and link them to each other. However, this requires some coordinated efforts on the level of, for instance, the European Union.

As regards the created dataset of real industrial energy efficiency improvement measures, its potential for overcoming barriers, and improving energy efficiency in industry, further research is suggested. However, a not too optimistic ex-ante evaluation of integrating the dataset with, e.g., current industrial energy audit programs, would be that up to 1 % further improved energy efficiency, directly or indirectly, would be reached. Supported by previous evaluations of energy audit programs and voluntary agreements (VAs), showing that measures are often found among the the support processes, the dataset would help reduce the risk for auditors and companies to suggest and implement measures that have already been implemented before (Fleiter et al. 2012; Stenqvist and Nilsson 2012). Case study research has already

been conducted by Cagno et al. (2010) using the IAC database in European conditions. We suggest future research to be conducted on the quantitative and qualitative gains of using a database within a national industrial energy efficiency program. A future case study should explicitly calculate how many more measures that are proposed and implemented thanks to the database in an industrial energy efficiency policy program, compared with “business as usual,” i.e., a program without a database, and also how much energy that could be further saved.

**Acknowledgements** This research was carried out within the DEFRAM project. We kindly thank the Swedish Energy Agency for financial support of DEFRAM, and we would in particular like to thank the project officer, Lara Kruse, for all the support given to this project. We would also like to express our thanks to Prof. Michael R. Muller, Rutgers University, who kindly allowed us to reuse the IAC database within our project. We additionally thank Robin Keskiärrkkä for developing the demonstration interface and providing the screenshots used in this paper for illustrating it. Finally, we give acknowledgement to Svetlana Paramonova, Danica Ilic, Kaihong Sun, Joel Forsberg, and Joel sterqvist for the processing and quality controlling of the EKC-dataset.

## References

- Anderson, S.T., & Newell, R.G. (2004). Information programs for technology adoption: the case of energy-efficiency audits. *Resource and Energy Economics*, 26 (1), 27–50.
- Berners-Lee, T. (2009). Linked data. <http://www.w3.org/DesignIssues/LinkedData.html>.
- Bizer, C. (2009). The emerging web of linked data. *IEEE Intelligent Systems*, 24(5), 87–92.
- Bleyl, J.W., & Eikmeier, B. (2009). *Energy Contracting: How much can it Contribute to Energy Efficiency in the Residential Sector? Transaction and Life Cycle Cost Analyses*.
- Cagno, E., Trucco, P., Trianni, A., & Sala, G. (2010). Quick-e-scan: a methodology for the energy scan of SMEs. *Energy*, 35(5), 1916–1926.
- Davis, C. (2012). Making sense of open data - from raw data to actionable insight. PhD thesis, Delft University of Technology.
- Djuric Ilic, D., & Trygg, L. (2014). Economic and environmental benefits of converting industrial processes to district heating. *Energy Conversion and Management*, 87, 305–317.
- Ehrlich, P.R., & Holdren, J.P. (1971). Impact of population growth. *Science*, 171(3977), 1212–1217.
- Enrico, C., & Trianni, A. (2014). Evaluating the barriers to specific industrial energy efficiency measures: an exploratory study in small and medium-sized enterprises. *Journal of Cleaner Production*, 82(1), 70–83.

- Fleiter, T., Gruber, E., Eichhammer, W., & Worrell, E. (2012). The German energy audit program for firms—a cost-effective way to improve energy efficiency *Energy Efficiency*, 5(4), 447–469.
- Franck, P., Åsblad, A., Thollander, P., & Nilsson, P.E. (2014). Teknikupphandling inom industrisektorn [technology procurement within the industrial sector]. [in Swedish], CIT,(Chalmers Industrial Management), Gothenburg.
- IAC (2014). URL <http://iac.rutgers.edu/database/>.
- IPCC (2007). Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Summary for Policymakers., retrieved October 8, 2007, from: <http://www.ipcc.ch/SPM040507.pdf>.
- Karlsson, M., & Thollander, P. (2012). Rohdin P (2012) Evaluation of the Swedish national energy audit program for small and medium sized industries. In *Proceedings of the ECEEE Industry Summer Study, 11-14 September*.
- Muller, M.R., & Kasten, D.J. (2007). Industrial Assessment Center Recommendation Code - The ARC., [http://iac.rutgers.edu/redirect.php?rf=ARC\\_list.9.1](http://iac.rutgers.edu/redirect.php?rf=ARC_list.9.1), last retrieved, February 27.
- Odyssee-Mure (2014). URL <http://www.odyssee-mure.eu/>.
- Ostertag, K. (1999). Transaction costs of raising energy efficiency. In *Proceedings of the IEA international workshop on technologies to reduce greenhouse gas emissions: engineering-economic analyses of conserved energy and carbon, 5-7 May 1999, Washington, D.C.*
- Söderström, M., Carlsson, P., & Hall, O. (1994). ENHET-SPROCESSER - Ett sätt att strukturera industrins energianvändning. Tech. rep., Linköping University and Syd-kraft Konsult.
- Sorrell, S., O'Malley, E., Schleich, J., & Scott, S. (2004). The economics of energy efficiency: barriers to cost-effective investment, vol 978 1 84064 889 8. Edward Elgar, Cheltenham.
- Spice 3 (2014). URL [http://www.eaci-projects.eu/iee/page/Page.jsp?op=project\\_detail&prid=2622](http://www.eaci-projects.eu/iee/page/Page.jsp?op=project_detail&prid=2622).
- Stenqvist, C., & Nilsson, L. J. (2012). Energy efficiency in energy-intensive industries—an evaluation of the Swedish voluntary agreement PFE. *Energy Efficiency*, 5(2), 225–241.
- Swedish Energy Agency (2010). Energy in Sweden 2010.Swedish Energy Agency Publication Department, Eskilstuna.
- Swerea Swecast (2014). URL <http://enig.se/energiinformation/energiledning-light/>.
- The OWL Working Group (2009). Owl 2 - web ontology language primer (second edition) - w3c recommendation. <http://www.w3.org/TR/owl2-primer/>, last retrieved February 27, 2014.
- The RDF Core Working Group (2004). RDF Primer—W3C Recommendation. <http://www.w3.org/TR/rdf-primer/>, last retrieved February 25, 2014.
- The W3C SPARQL Working Group (2013). SPARQL 1.1 Overview - W3C Recommendation. <http://www.w3.org/TR/sparql11-overview/>, last retrieved February 27, 2014.
- Thollander, P., & Dotzauer, E. (2010). An energy efficiency program for Swedish industrial small- and medium-sized enterprises. *Journal of Cleaner Production*, 18(13), 1339–1346.
- Thollander, P., & Ottosson, M. (2008). An energy efficient Swedish pulp and paper industry—exploring barriers to and driving forces for cost-effective energy efficiency investments. *Energy Efficiency*, 1(1), 21–34. doi:10.1007/s12053-007-9001-7.
- Thollander, P., & Palm, J. (2013). *Improving Energy Efficiency in Industrial Energy Systems—An Interdisciplinary Perspective on Barriers, Energy Audits, Energy Management, Policies, and Programs*, vol 978-1-4471-4161-7. London: Springer.
- Thollander, P., Svensson, I.L., & Trygg, L. (2010). Analyzing variables for district heating collaborations between energy utilities and industries. *Energy*, 35(9), 3649–3656.
- Thollander, P. et al. (2014). International study on energy end-use data among industrial smes and energy-end-use efficiency improvement opportunities. Submitted for publication in *Journal of Cleaner Production*.
- W3C Semantic Web Deployment Working Group (2012). SKOS Simple Knowledge Organization System - Home Page. <http://www.w3.org/2004/02/skos/>.
- Waide, P., & Brunner, C. U. (2011). *Energy-efficiency policy opportunities for electric motor-driven systems*. Tech. Rep. 2011/07, International Energy Agency, Paris.