

Adopting Semantic Technologies for Effective Corporate Transparency

Maria Mora-Rodriguez¹(✉), Ghislain Auguste Ateazing², and Chris Preist¹

¹ System Centre, University of Bristol, Bristol, UK
{[maria.mora](mailto:maria.mora@bristol.ac.uk), [chris.preist](mailto:chris.preist@bristol.ac.uk)}@bristol.ac.uk

² Mondeca S.A, 35 boulevard de Strasbourg, Paris, France
ghislain.ateazing@mondeca.com

Abstract. A new transparency model with more and better corporate data is necessary to promote sustainable economic growth. In particular, there is a need to link factors regarding non-financial performance of corporations - such as social and environmental impacts, both positive and negative - into decision-making processes of investors and other stakeholders. To do this, we need to develop better ways to access and analyse corporate social, environmental and financial performance information, and to link together insights from these different sources. Such sources are already on the web in non-structured and structured data formats, a big part of them in XBRL (Extensible Business Reporting Language). This study is about promoting solutions to drive effective transparency for a sustainable economy, given the current adoption of XBRL, and the new opportunities that Linked Data can offer. We present (1) a methodology to formalise XBRL as RDF using Linked data principles and (2) demonstrate its usefulness through a use case connecting and making the data accessible.

Keywords: XBRL · Transparency · Linked data · Open government data · Interoperability

1 Introduction

Transparency is increasingly used as a means for holding to account organisations with power, both in the public and private sector. In this paper, we focus primarily on the latter - the role of transparency and open data to promote good governance in the private sector, and trust between the private sector and its diverse stakeholders. The main tool for this is corporate reporting - the self-disclosure of information by a company in a well-defined and regular way to a set of stakeholders - primarily, though not exclusively, existing and potential investors. Such stakeholders need to satisfy themselves as to the financial performance and good governance of the company they invest in. Many governments require such reports to be in specific formats, and increasingly are making them

publically available through open data initiatives – such as the EDGAR¹ program of the U.S. Security and Exchange Commission and the data repository of the Spanish Security Exchange Commission (CNMV)².

Such data is often submitted and made available in XBRL (Extensible Business Reporting Language), an XML format adopted to make corporate data more standardised and exchangeable. XBRL is currently in use in more than 60 countries, implemented by over 100 regulators covering some 10 million companies worldwide.

In addition to mandatory reporting, there are a number of voluntary initiatives encouraging corporations to disclose information regarding their performance in areas of economic, social and environmental impact. These include the CDP (Carbon Disclosure Project), GRI (Global Reporting Initiative), SASB (Sustainability Accounting Standards Board) and the IIRC (International Integrated Reporting Council). These are often international non-governmental organisations, with representatives of both corporations and other stakeholders including investors, academics, environmental NGOs and policymakers in their governance structures. Some of these, are promoting the disclosure and use of their data through XBRL, in a similar way to governmental open data initiatives.

Such initiatives encourage engagement with this data: Data journalists can investigate corporate behaviour; investors can integrate future risks associated with factors such as climate change in their assessment of companies; companies can benchmark themselves against others; academics can explore wider trends and correlations in financial performance and non-financial behaviours.

However, such transparency alone is not enough to support companies and their stakeholder's decisions and actions. There are several barriers such as:

- The lack of easily accessible corporate data to quickly and accurately inform the management about issues pertinent during decision-making-processes.
- The unfamiliarity of how sustainable aspects have an impact on financial outcomes and vice versa.
- Inadequate levels of integration of financial and non-financial information within the internal performance, strategy and operational frameworks of an organisation.
- A dearth of consistency, comparability, reliability and clarity of climate change information emerging from organisations globally. Standardisation and mainstreaming of disclosures needs to be facilitated.

The exposure of XBRL reports as open data is a first step to overcome some of these barriers. XBRL brings access to standardised data in an open format about corporate financial, sustainability and environmental performance, in independent silos. However, XBRL offers limited interconnection between them, In particular, XBRL exhibits the following weaknesses:

¹ <https://www.sec.gov/xbrl/site/xbrl.shtml>.

² <http://www.cnmv.es/ipps/Default.aspx>.

- It is primarily structured around documents and entities, rather than data making links between data elements difficult.
- The same data structures can be modelled in different ways, generating different technical implementations.

The Open Information Model (OIM) is being developed by the XBRL community in an effort to increase the adoption of existing XBRL data by reducing the heterogeneity of the XBRL reports generated by different modelling practices and enabling a better integration with Information Systems. OIM aims to facilitate the serialisation of XBRL reports in CSV, JSON and XML formats. In this paper, we use the XBRL Open Information Model as a template for extracting linked open data from XBRL documents, and so allowing them to be combined more easily and to be used alongside other open data sources to enrich analysis by stakeholders. We demonstrate this by reasoning with data from the Spanish Security and Exchange Commission (CNMV) alongside data from the CDP.

Our paper makes a contribution in the field of the Semantics and Corporate Transparency, by more effectively integrating data sources using the most common corporate standard - XBRL - into the linked data environment. We show that this in turn makes contributions in the following areas:

- The generation and integration of Linked Open Government data made available from government sources (CNMV) in a different format (XBRL).
- Provenance and Accountability of companies, showing how financial and non-financial data from different sources can be linked using our approach to hold companies more environmentally accountable.
- Trust, Data Traceability and Fact Checking of corporate data sources, enabled by using the proposed ontology together with SPARQL to facilitate cross-checking of corporate data with alternate sources.

This paper is organised as follows. In Sects. 2 and 3, we provide background on XBRL and prior work connecting XBRL with Linked Data formats. Section 4 presents a methodology to transform XBRL to RDF following Linked Data principles and the Open Information Model³. We describe a case study implementing our proposal, using financial and environmental data in XBRL format from the Spanish companies Repsol S.A and Amadeus IT Group, published by the CNMV and CDP respectively. Having XBRL data in RDF format, we proceed to link with others data sources from LOD using LIMES⁴ as Link Discovery Framework. Then, we make the RDF data accessible and queryable via SPARQL endpoint, and we proceed to evaluate some queries to show the potential benefits for data users. Section 6 presents and discusses the results. The paper closes with conclusions and lessons learned which, if addressed by the XBRL and Linked data community, would further promote transparency.

³ <http://www.xbrl.org/Specification/oim/PWD-2016-01-13/oim-PWD-2016-01-13.html>.

⁴ <http://aksw.org/Projects/LIMES.html>.

2 XBRL Fundamentals

XBRL aims to overcome the limitations in traditional and mainly paper-based disclosures [7]. Through data standardisation in an open digital format, XBRL can help enhance data quality and data analysis through:

- An open mechanism to represent contextualised business facts under defined business requirements (presentation, period, legal references, calculation) and data quality;
- Enabling data-driven decision management;
- Improved accessibility and integration of the information to any application or management process;
- Standardised validation and comparability of information.

For financial disclosure regimes like International Financial Reporting Standards (IFRS) and US Generally Accepted Accounting Principles (US GAAP) or non-financial disclosure regimes like the CDP and GRI, and its corresponding statements and reports, a single XBRL taxonomy is created. The taxonomy is where the rules and data definitions are organised. It is comprised of a set of elements (i.e., Key Performance Indicators and narratives) and all the presentation, calculation, labels in different languages, and standard logic rules (linkbases) which provide semantic meaning. The taxonomy also includes mechanisms for defining reporting requirements in a multi-lingual context and allowing filers to use their language of choice while allowing consumers to review it in their own.

Once created, the XBRL taxonomy is published online. Then, for a given firm, software can be used to create an XBRL instance (the report itself), containing facts and figures for a certain period (Fig. 1). The XBRL instance can be checked against the taxonomy by all parties (reporting entity, a regulator, or even the public) in order to guarantee its data quality and reliability, as the taxonomy contains data quality checks that any XBRL engine can validate. The validation rules supported in XBRL allow a good level of data quality, from basic rules to validate data types (number, text, precision), to more complex rules relating to elements that have been disclosed. For example, rules can be implemented to check if a breakdown of carbon emissions is equal or not to the total emissions reported, or a CO₂ intensity figure (tCO₂/revenue) is actually in line with revenue and emissions figures reported.

The XBRL core specification (XBRL 2.1) has evolved since its creation in 2001 to enrich its dimensional data representation and validation rules. The XBRL Dimensions 1.0⁵ in 2006 and Formulas 1.0⁶ in 2009 provide optional incremental syntax to the core specification. Dimensions 1.0 enriches rules and procedures for constructing dimensional taxonomies and therefore instance documents. Taxonomies using Dimensions specification can define new dimensional

⁵ <http://www.xbrl.org/specification/dimensions/rec-2012-01-25/dimensions-rec-2006-09-18+corrected-errata-2012-01-25-clean.html>.

⁶ <http://www.xbrl.org/wgn/xbrl-formula-overview/pwd-2011-12-21/xbrl-formula-overview-wgn-pwd-2011-12-21.html>.

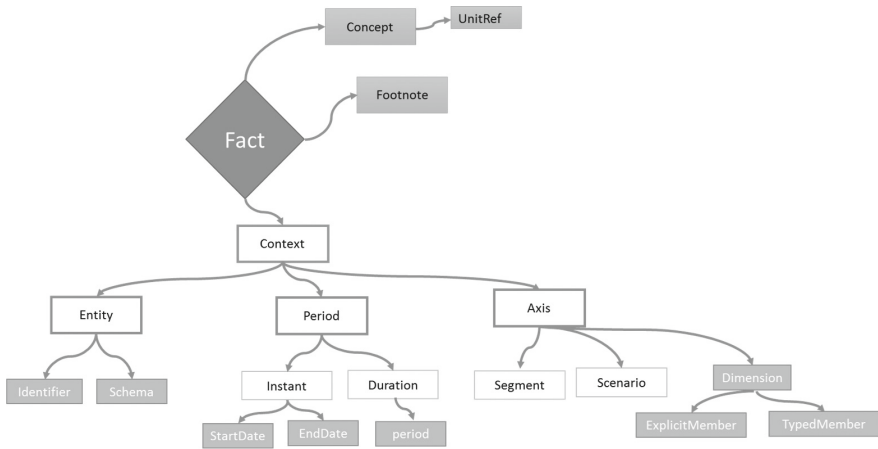


Fig. 1. Fact representation in XBRL

contexts, specifying valid values (“domains”) for dimensions, using a mechanism called hypercube to define which dimensions apply to which business concept. There are two types of dimensions:

- **Explicit dimensions:** These have a fixed number of dimension members. For example, in a two-dimensional table, the number of row and columns are known.
- **Typed dimensions:** the number of dimension members is unknown. For example, in a two-dimensional table, it means that the number of columns is known, but the number of rows depends on user reporting needs.

However, the Dimensions specification in XBRL 2.1 has certain limitations. It does not fully support calculation rules (defined in calculation linkbase); calculations cannot be executed across different contexts. In other words, in a simple two-dimensional table, calculations can be executed by columns and not by rows. In part, these limitations, and the need to have strong validation capabilities in XBRL resulted in the Formulas 1.0 specification in 2009. This module enhances the XBRL validation capabilities, using XPath to validate instances and to calculate new XBRL facts.

XBRL has evolved to support both global regulatory environments and emerging domains such as sustainability reporting. The flexibility needed to do this has resulted in difficulties with regard to standardisation. Primarily, this is because of the diversity of technical implementations produced by different modelling practices during the taxonomy development phase. Though XBRL is a standard, it offers different ways to model data structures. For example, tables can be represented in XBRL using tuples or dimensions. This is true of the two taxonomies we use in our work later in the paper: The CNMV taxonomy represents financial facts as items and tuples and only makes use of the XBRL 2.1 core specification. Items are facts holding a simple value represented by a

single XML element with the value as its content and period and information about reporting entity as a context attribute. An example item could be equity in the last quarter. Tuples are facts holding multiple values, and they are represented by a single XML element containing items or other tuples. For example, preferred stock is always defined by the combination of different stocks. Thus, preferred stock is a tuple defined by two items: the preferred stock-nominal value and the preferred stock-shares authorised. The CDP taxonomy, on the other hand, represents (environmental) facts as simple items and dimensional structures and makes use of XBRL 2.1 core and Dimensions 1.0 specifications. Dimensional structures are represented as items where context attributes also include dimensional XML elements.

The new Open Information Model (OIM) specification proposes an independent XBRL model to represent XBRL business facts; it focuses on XBRL instance documents instead of taxonomy definitions. This overcomes the taxonomy modelling difficulties and keeps the value and context of the data. However, the semantic richness of the taxonomy disappears, such as advanced validation rules, human labels in different languages and how the data should be presented.

In our study, we propose an ontology based on OIM specification to transform XBRL data to RDF, using as a case of study financial and environmental company data from the CNMV and CDP.

3 Related Work

Previous efforts to make XBRL data more interoperable with other data sources and formats have used RDF and OWL ontologies, as well as linking and publishing solutions. The majority of these base their examples on transforming financial XBRL taxonomies models into RDF from well-known open government data initiatives, such as XBRL filings available from the SEC's EDGAR program. However, none of these studies covers the full XBRL specifications including XBRL 2.1 and Dimensions 1.0. In other words, these studies do not offer a general solution to convert any XBRL report to RDF. For example, Garcia and Gil [2], propose a solution to transform XBRL filings available from the EDGAR program to RDF. Their approach is generic to the XBRL 2.1 specification: simple items, scenarios, segments and tuples data structures. They use US-GAAP reports from 2006 as a case study, which do not use Dimensions 1.0 specification. On the other hand, Kampgen *et al.* [5], propose RDF Data Cube Vocabulary to model XBRL reports as a multidimensional dataset. They exemplified their methods by using 2009 and 2011 US-GAAP reports, whose taxonomy uses XBRL 2.1 and Dimensions 1.0 specification. However, it is unclear how that solution can be generic to other dimensional taxonomies and how the ontology proposed covers tuples, simple items and contextual information modelled with scenarios and segments. There is an experimental initiative, called the Edgar Linked Data wrapper⁷, that provides access to XBRL filings from the SEC as Linked Data. The approach is to publish US-GAAP taxonomies

⁷ <http://edgarwrap.ontologycentral.com/>.

into RDF as vocabularies. In fact, each new US-GAAP taxonomy version means a new semantic vocabulary. This represents a solution to convert US-GAAP reports into Linked data, but it is not a solution for any other type of XBRL reports, such as CNMV reports. Closer to the sustainability domain, Madlberger *et al.* [6] presented an ontology-based approach using GRI-XBRL taxonomy to build a Corporate Sustainability ontology. However, the result is content-based approach instead of metadata conversion, meaning that the solution proposed is not generic to transform any XBRL report to RDF, only GRI reports to RDF.

Authors agree that there are some limitations when representing XBRL data in RDF graphs and as Linked data, due to the lack of formal semantics and inference mechanisms, and difficulties to find correspondences with well-known vocabularies (SKOS, FOAF, etc.). Furthermore, complete architectures for evolving information systems enabling a better financial data integration using Linked data are proposed in [4,5]. Basically, these solutions integrate XBRL financial data with DBpedia and Yahoo!Finance Web API. For the purpose of this study, we also consider their requirements necessary to boost effective corporate transparency:

- to break the barriers which hold XBRL data in isolated data silos of information and vendor lock-in of proprietary XBRL tools,
- to reach a better level of data coverage and data quality and
- to facilitate a comprehensive picture of company performance.

We distinguish our study from previous work by proposing a solution to enhance an effective corporate transparency, increasing the adoption of financial and non-financial data and generating impact on decisions. Our central thesis is that in order to create that impact, two components are necessary: (1) Foster interoperability across economic, social and environmental data published in XBRL format and others; and (2) better integration of these combined data in information systems that are part of the decision-making processes of companies, their stakeholders, regulators and supervisory entities. In order to turn corporate data into valuable information for decision-makings, we focus on the following tasks:

- A generic ontology to transform any XBRL report into Linked data.
- Interlinking with existing data available in the LOD cloud.
- Data publication via SPARQL query endpoint.
- Enabling data contextualization, cross-data-source analyses and data accuracy.

4 Transforming XBRL into Linked Data

4.1 Lightweight Vocabulary for XBRL (XBRL-L)

We develop a lightweight ontology using the Web Ontology Language (OWL) based on XBRL standard (XBRL 2.1 and Dimensions 1.0). The goal of implementing a lightweight vocabulary for XBRL is threefold: (1) Easy identification

of the key concepts of the XBRL standard; (2) Reuse of existing vocabularies to describe XBRL datasets; (3) Enrichment and linking of data with relevant ones in the Linked Open Data cloud.

Unlike previous efforts, we base our ontology proposal on OIM, which is a syntax-independent model of the content of an XBRL report instead of taxonomy definition. OIM defines 4 components:

Namespace: Representation of XML namespace prefixes.

DTS Reference: Reference to XML documents and schema linked to an XBRL report.

Report: Top-level component that encapsulates the data of an XBRL report.

Fact: Representation of a business fact in an XBRL report. As explained in section one, a fact can be a simple item, a tuple fact or a dimension. All facts have the following common properties (id, aspects and footnotes). Id is a unique identifier; Aspects are properties which represent:

- the entity and period which a business fact is referred to (oim:entity and oim:period),
- the unit of measure, such as “USD” and “MWh” (oim:unit),
- the reporting item (oim:concept),
- tuples definition, which represents a grouping container for other facts.
- dimensional structure, axis and members.
- the footnotes of the fact (oim:footnotes)

Our XBRL-Lightweight (XBRL-L) ontology is composed of 15 classes, 12 object properties and 12 data properties, with DL expressivity: ALC(D). The ontology follows best practice in the semantic web by reusing existing ontologies to improve data interoperability [1], through the Linked Open Vocabularies initiative (LOV⁸).

We provide a hierarchical structure following OIM, mapping XBRL components to Semantic web vocabularies. The class Report is a subclass of schema:Report⁹. The class Fact is used to represent XBRL business facts (items, tuples and dimensions) that in turn refer to entity (hasEntity), concept, period, scenario, value and footnote modelled as object properties. In the case of numeric and currency values, the number of decimals and unit type are also represented. These properties are enough to represent a simple XBRL item.

As required from XBRL, a fact can hold multiple values in the form of tuples and dimensional structures. For that purpose, we defined the has Tuple and hasDimension properties as part of the Fact class. They point out to Tuple and Dimension classes respectively. Tuple class is composed by concept, and hasTuple properties. The latter means a tuple can be embedded as part of another tuple, defining the context of the main item. Dimension class is composed of Axis and Member properties, representing the axis and member per axis which define the context of the main item. We decided not to differentiate whether the axis is explicit or typed, as our model is focused on instance documents instead of

⁸ <https://lov.okfn.org/>.

⁹ <http://schema.org>.

taxonomy definition. It means, that as we are working with XBRL reports, the dimensional members are defined.

Unlike the OIM model, we decided to define Entity as a class instead of an object property of the class Fact that links to the schema identifier of the entity that is part of the XBRL report. In that way, we extend the class `rov:RegisteredOrganization`¹⁰, (1) keeping the correspondence with well-known vocabularies and (2) allowing the full information of the reporting firm facilitating the discovering link process. Figure 2 shows classes and relationships defined in the XBRL ontology, which is available at <https://w3id.org/vocab/xbrll>.

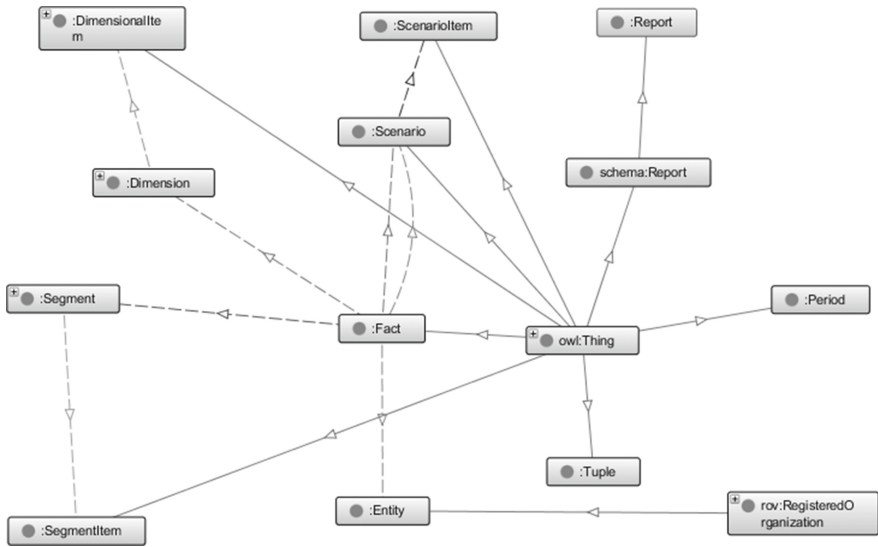


Fig. 2. Classes and relationships XBRL ontology

4.2 From XBRL Data to Linked Data

As a next step, we demonstrate how XBRL data can be mapped to the ontology and how it can be published using Linked data principles. We demonstrate our method using financial and environmental XBRL data from the Spanish companies Repsol and Amadeus IT Group, chosen as ones which are published in both CNMV and CDP¹¹. For the transformation process, we developed a script in Python (<https://goo.gl/VqgJQZ>) to transform the XBRL reports using JSON files generated by the Arelle¹² open source platform. Note that the JSON data used as our input is generated according to OIM.

¹⁰ <http://www.w3.org/ns/regorg>.

¹¹ The CDP data used is publically available, but not yet in XBRL format. This is currently only available internally to CDP, and made available to this project.

¹² <http://arelle.org/>.

A simple fact in XBRL (Fig. 3), is the representation of a concept (cdp:IntroductionCompany) and its value, where the context consists of the period, unit and information about the reporting company (entity).

A tuple fact in XBRL (Fig. 4) represents facts with multiple values. In this case, the concepts ifrs-gp:IntangibleAssetsNet and ifrs-gp:GoodwillNet are the elements that compose the tuple `ipp-gen:BalanceIndividual`. The CNMV allows reporting companies to use `xbkli:scenario` element to determine if the value is part of an individual or consolidated financial statement. Hence in CNMV reports we find the same tuple hierarchy and concepts linked to two different contexts: consolidated and individual. Currently, the scenario and segment elements, in a non-dimensional domain, such as the CNMV reports, are not considered either by OIM and Arelle when transforming XBRL into JSON¹³. Our ontology considers both.

<pre><http://data.mondeca.com/id/fact/f001> a nsl:Fact ; nsl:hasEntity <http://data.mondeca.com/id/entity/000084 7838> ; nsl:concept cdp:IntroductionCompany; nsl:period [nsl:startPeriod "2015- 01-01"^^xsd:date ; nsl:endPeriod "2015-12- 31"^^xsd:date] ; nsl:value "Repsol is international integrated Oil and Gas Company whose main activity consists of the upstream and downstream business." ;</pre>	<pre><xbkli:context id="ctx_00"> <xbkli:entity> <xbkli:identifier scheme="http://www.cdp.net/CIK">0000847838</xbkli:identifie r> </xbkli:entity> <xbkli:period> <xbkli:startDate>2015-01-01</xbkli:startDate> <xbkli:endDate>2015-12-31</xbkli:endDate> </xbkli:period> </xbkli:context> <cdp:IntroductionCompany contextRef="ctx_00">Repsol is international integrated Oil and Gas Company whose main activity consists of the upstream and downstream business.<cdp:IntroductionCompany></pre>
--	---

Fig. 3. RDF representation of a simple XBRL fact

<pre><http://data.mondeca.com/id/fact/f0004> a nsl:Fact ; nsl:concept ifrs-gp:IntangibleAssetsNet; nsl:entity <http://data.mondeca.com/id/entity/A7837472>; nsl:period [nsl:instant "2015-12-31"^^xsd:date] ; nsl:hasTuple <http://data.mondeca.com/id/tuple/t89493> ; nsl:unitRef < http://dbpedia.org/resource/EUR> ; nsl:value 94588000.0 ; nsl:decimals 0 . nsl:hasScenario <http://data.mondeca.com/id/scenario/s0001> . <http://data.mondeca.com/id/fact/f0004> a nsl:Fact ; nsl:concept ifrs-gp:GoodwillNet ; nsl:entity <http://data.mondeca.com/id/entity/A7837472>; nsl:period [nsl:instant "2015-12-31"^^xsd:date] ; nsl:hasTuple <http://data.mondeca.com/id/tuple/t89493> ; nsl:unitRef < http://dbpedia.org/resource/EUR>; nsl:value 350.0 ; nsl:decimals 0 . nsl:hasScenario <http://data.mondeca.com/id/scenario/s0001> . <http://data.mondeca.com/id/scenario/s0001> a nsl:Scenario nsl:scenarioItem [nsl:concept ipp:Modelo; nsl:value "GE"] nsl:scenarioItem [nsl:concept ipp:Apartado; nsl:value "Individual"] <http://data.mondeca.com/id/tuple/t89493> a nsl:Tuple ; nsl:concept ipp-gen:BalanceIndividual; nsl:hasTuple <http://data.mondeca.com/id/tuple/t89494></pre>	<pre><xbkli:context id="S22015_A-78374725_ici"> <xbkli:entity> <xbkli:identifier scheme="http://www.cnmv.es/xbrl/ipp/A- 78374725">REPSOL, S.A.</xbkli:identifier> </xbkli:entity> <xbkli:period> <xbkli:instant>2015-12-31</xbkli:instant> </xbkli:period> <xbkli:scenario> <ipp:Modelo>GE</ipp:Modelo> <ipp:Apartado>Individual</ipp:Apartado> </xbkli:scenario> </xbkli:context> <xbkli:unit id="euro"> <xbkli:measure>iso4217:EUR</xbkli:measure> </xbkli:unit> <ipp-gen:InformacionFinancieraSeleccionada> <ipp-gen:BalanceIndividual> <ifrs-gp:IntangibleAssetsNet decimals="0" contextRef="S22015_A-78374725_ici" unitRef="euro">94588000</ifrs- gp:IntangibleAssetsNet> <ifrs-gp:GoodwillNet decimals="0" contextRef="S22015_A-78374725_ici" unitRef="euro">350</ifrs-gp:GoodwillNet> </ipp-gen:BalanceIndividual> </ipp-gen:InformacionFinancieraSeleccionada></pre>
--	--

Fig. 4. RDF representation of a tuple fact

¹³ The lack of segment and scenario representation in a non-dimensional domain, was informed to the OIM working group and Arelle’s authors.

Dimensional facts in XBRL (Fig. 5), can also represent multiple values. For example, the concepts `cdp:EmissionValueGross` and `cdp:Scope` are linked to the same Axis `cdp:TotalEmissionDataAxis` and related member(`cdp:GreenhouseInventoryBoundariesID`). This structure allows disclosing the total Emissions gross values per type of scope (Scope 1, Scope 2 location-based, Scope 2 market-based and Scope 3). Here the value of 21068516 CO₂e corresponds to the Scope 1 emissions.

<pre><http://data.mondeca.com/id/fact/f0005> a ns1:Fact ; ns1:concept cdp:EmissionValueGrossCO2e ns1:entity <http://data.mondeca.com/id/entity/00008478 38> ; ns1:period [ns1:startPeriod "2015-01- 01"^^xsd:date; ns1:endPeriod "2015-12- 31"^^xsd:date] ; ns1:hasDimension <http://data.mondeca.com/id/dimension/d0001 > ; ns1:unitRef <http://dbpedia.org/page/Carbon_dioxide_equ ivalent>; ns1:value 21068516; ns1:decimals 0 . . <http://data.mondeca.com/id/dimension/d0001 > a ns1:Dimension ; ns1:dimensionItem [ns1:axis TotalEmissionDataAxis; ns1:value "id01"]</pre>	<pre><xbrli:context id="ctx_8_2_2015_id01"> <xbrli:entity> <xbrli:identifier scheme="http://www.cdp.net/CIK">0000847838</xbrli:identifier> <xbrli:segment> <xbrldi:typedMember dimension="cdp:TotalEmissionDataAxis"> <cdp:GreenhouseInventoryBoundariesID>id01</cdp:GreenhouseInvento ryBoundariesID> </xbrldi:typedMember> </xbrli:segment> </xbrli:entity> <xbrli:period> <xbrli:startDate>2015-01-01</xbrli:startDate> <xbrli:endDate>2015-12-31</xbrli:endDate> </xbrli:period> </xbrli:context> <xbrli:unit id="CO2e"> <xbrli:measure>cdp:CO2e</xbrli:measure> </xbrli:unit> <cdp:EmissionValueGrossCO2e decimals="0" contextRef="ctx_8_2_2015_id01" unitRef="CO2e">21068516</cdp:EmissionValueGrossCO2e> <cdp:Scope contextRef="ctx_8_2_2015_id01">cdp- enum:Scope1</cdp:Scope></pre>
---	--

Fig. 5. RDF representation of a dimensional fact

During the transformation of the CNMV report, we found certain XBRL elements with content about persons and activities, which belongs to the imported XBRL taxonomy called Data of General Identification (DGI)¹⁴. We map these using Friend Of a Friend (FOAF) vocabulary¹⁵ (Fig. 6), in line with best practice of reusing existing vocabularies in specific contexts to increase the level of interoperability.

As in XML [8], XBRL Namespaces specifications do not need to reference a real location, just be unique. However, in RDF, the namespace URI must identify the location of the schemas. As certain XBRL namespaces from the CNMV reports were not valid we had to store the schemas in our server and point out the namespaces to real locations. In many cases, we decided to map the XBRL units and currencies to well-known DBpedia links, connecting related data that were not previously linked. For example:

`<xbrli:measure>iso4217:EUR</xbrli:measure>` to <http://dbpedia.org/resource/EUR>

¹⁴ https://joinup.ec.europa.eu/asset/data_of_general_identification/home.

¹⁵ <http://xmlns.com/foaf/spec/>.

<pre><http://data.mondeca.com/id/xbrl/person/01606701J> a ns5:Person ; ns7:identifier "01606701J" ; ns6:role "Consejero" ; ns5:name "D. Javier Echenique Landiribar" ns2:codeID "CIF" .</pre>	<pre><ipp-com:DetallePersonasResponsables> <dgi-est-gen:PersonName> <dgi-est-gen:FormattedName contextRef="S12015_A-78374725_ici">D. Ángel Durández Adeva</dgi-est-gen:FormattedName> </dgi-est-gen:PersonName> <dgi-est-gen:Identifier> <dgi-est-gen:IdentifierCode> <dgi-lc-es:Xcode_IDC.CIF contextRef="S12015_A-78374725_ici">CIF</dgi-lc-es:Xcode_IDC.CIF> <dgi-est-gen:IdentifierValue contextRef="S12015_A-78374725_ici">01606701J</dgi-est-gen:IdentifierValue> </dgi-est-gen:IdentifierCode> </dgi-est-gen:Identifier> <dgi-est-gen:PositionType contextRef="S12015_A-78374725_ici">Consejero</dgi-est-gen:PositionType> </ipp-com:DetallePersonasResponsables></pre>
---	--

Fig. 6. DGI representation in RDF using FOAF vocabulary

`<xbrli:measure>cdp:CO2e</xbrli:measure>` to http://dbpedia.org/page/Carbon_dioxide_equivalent

As XML, an XBRL document forms a tree structure ready to be consumed as a full report [3]. The move to data consumption requires the use of dereferenceable URIs to denote facts in a unique way, keeping its context. We use the following URI conventions to denote related facts and classes:

Fact: `http://data.mondeca.com/id/fact/f[0-9]*` -> <http://data.mondeca.com/id/fact/f88557>

Entity: `http://data.mondeca.com/id/entity/(company identifier)` -> <http://data.mondeca.com/id/entity/A-78374725>

Through the dereferenceable URIs, facts can be visualised using open source tools like LodLive¹⁶. We provide an example here: <https://goo.gl/iFVE0B>.

4.3 Linking XBRL Data to Other Data

If transparency is to be enabled, it is very important to convert the independent XBRL silos of information into pieces connected with existing Linked Data sources available on the web.

For that purpose, we use LIMES, which is a tool that allows detecting similar Linked datasets. LIMES works specifying the search criteria and the target endpoint to search in. Our search criteria is the company name contained in the Entity fact from the generated RDF. The DBpedia endpoint is the target source that we choose to gather the links, restricting the search by `sch:Organization`. LIMES requires a metric and acceptance condition setting a threshold value. We use the trigrams metric offered by LIMES to mapping correspondences between the `ns6:legalName` of our local RDFs and the `sch:Organization` from DBpedia. For the purpose of this paper, we only accept results with a minimum of 0.90 level acceptance. The final results were the following URLs <http://dbpedia.org/resource/Repsol> and http://dbpedia.org/page/Amadeus_IT_Group, included as a SameAs relationship in our local RDFs files.

¹⁶ <http://en.lodlive.it/>.

5 Validation

For validation purposes, we run queries against the final ontology generated, evaluating its quality and accuracy by checking whether they contain enough information to cover three goals to promote effective transparency:

Data coverage: through better data contextualization.

Better data analysis: enabling cross-data-source analyses.

Data accuracy: facilitating data cross-checking contained in different sources.

For that, firstly we implemented an endpoint¹⁷ using Apache Jena Fuseki¹⁸, available here: <http://data.mondeca.com/dataset.html?tab=query&ds=/xbrl-data>. We used SPARQL (Simple Protocol and RDF Query language) because it allows us to express queries across diverse data. We conduct three queries with each of the two companies (Repsol and Amadeus IT Group) data. We illustrate each query below with one of the companies. Full results are available at the SPARQL queries provided.

Goal 1. Data Coverage Using DBpedia

Question: What is the context of the company Repsol?

Data: Abstract, subsidiary and industry.

SPARQL query: <https://goo.gl/lf8ydG>

Output: presented in Table 1.

Table 1. Data coverage: information about the context of Repsol S.A

URL	Abstract	Subsidiary	Industry
http://dbpedia.org/resource/Repsol	Repsol S.A. is an integrated global energy company based in Madrid, Spain. It carries out upstream and downstream activities throughout the entire world. It has more than 24,000 employees worldwide	http://dbpedia.org/resource/Petronor	http://dbpedia.org/resource/Petroleum_industry

Goal 2. Cross Data Source Analysis Using CNMV and CDP Data

Question: What was the emission intensity of Repsol in 2015?

Data: Scope 1 emissions (CDP) divided by Consolidated sales (CNMV).

SPARQL query: <https://goo.gl/7bIE9m>

Output: presented in Table 2.

Table 2. Data analysis: emission intensity of Repsol S.A in 2015

CO2 Emissions	Consolidated sales	Emission intensity
21068516	3973700000	0.00053

¹⁷ <http://data.mondeca.com/xbrl-data/sparql>.

¹⁸ <https://jena.apache.org/>.

Goal 3. Data Accuracy Using DBPedia and CNMV Data

Question: How reliable is the equity figure presented in DBpedia?

Data: Equity (DBPedia) and equity (CNMV) in the year 2013.

SPARQL query: <https://goo.gl/LGb53s>

Output: presented in Table 3.

Table 3. Data consistency: reliability of equity figure presented in DBpedia

Entity name	Equity(DBPedia) dbo:equity	Equity(CNMV) ipp- gen:PatrimonioNetoNiif	Difference
Repsol S.A	2.792E10	27920000000	0.0001
Amadeus IT holding	€1,840.1 million@en	1840066000.0	-

6 Discussion

This study demonstrates that Linked data can be used to integrate financial and non-financial data and can facilitate transparency among diverse stakeholders. Our work does this by converting corporate XBRL reports into RDF and linking them to other relevant financial and non-financial data (e.g., environmental, DBpedia). A generic ontology to transform any XBRL report into Linked data has been proposed, along with ways to resolve the lack of formal correspondences with well-known vocabularies. This solution overcomes the XBRL challenges related to the diversity of technical implementations produced by different modelling practices, and so goes beyond prior related works.

We demonstrate that using linked data with well-adopted standards, such as XBRL, improves the interoperability and access to existing corporate datasets, as well as straightforward integration with related data in other formats. The validation exercise demonstrates that the solution proposed offers three benefits for data users: data coverage, better data analysis and data consistency. The results in Table 3 present an interesting point for discussion. It shows that the DBpedia data (dbo: equity) lacks context and numeric precision. For example, there is no year associated with the equity figure nor consistent use of datatype. Amadeus equity is a string €1,840.1 million while Repsol equity, which has the same tag (dbo:equity), is a number. The data from the CNMV in XBRL format does not have any of those problems.

Given these results, we believe that XBRL is a better format than RDF to standardise corporate information. However, it is less able to connect different data silos in various formats. For that, the publication of data using Linked data principles is the most appropriate solution.

This study proposes the combination of both solutions, XBRL and Linked data, to improve corporate transparency. Below, we enumerate a set of technical

requirements to apply on XBRL schemas, definitions and reports to converge towards a Linked data approach. Adoption of these best practices in XBRL modelling would enhance interoperability and transparency of corporate data.

- Use of common data structures must be encouraged in XBRL taxonomies. For example, taxonomies such as DGI to represent common corporate information such as company name, unique identification number, activities and sectors. This would not only enhance the interoperability between XBRL data from different taxonomies but also ease the mapping process with well-known vocabularies in the Linked data world.
- The use of namespaces notation that point out to real locations should be promoted in XBRL. This would ease the transformation of XBRL data into Linked data, facilitating better inference mechanisms.
- Using dereferenceable URIs to denote and identify XBRL facts provides a way to access and link relevant information to those objects across the web. It enables better interoperability between data in XBRL format and other data sources.
- Reusing existing RDF data on units and currencies already published in LOD brings more contextual information than the current ISO and XBRL units reference.
- Using tools like LIMES can help to increase the coverage of information by continuously integrating data sources.

We made all scripts and tools available to let academia and industry evaluate and contribute to this work.

7 Conclusions and Further Research

In this study, we show the role of Linked Data and XBRL in bringing new opportunities for effective transparency in corporate reporting. Linked data principles can encourage better corporate data publication and therefore data analysis, defining the interconnection across financial and non-financial data (such as sustainability data) and documents publicly available in open government data initiatives and voluntary reporting initiatives. XBRL enables a standard and accurate representation of corporate data with advanced validation rules. We present a solution to convert independent silos of XBRL data into interconnected pieces. Lessons learned during the process and benefits are presented. While our work demonstrates the potential of this approach, it would benefit from extension in the following ways; (i) incorporate data sources beyond environmental, financial and DBpedia; (ii) incorporate non public-domain data, and address associated data protection issues necessary to do this; (iii) consider scalability and performance issues in the transformations necessary. In future work, we intend to evaluate and integrate sustainability reporting in XBRL, such as GRI data, and extend the ontology proposed using RDF Data Cube. We believe this study encourages scholars, regulators, data publishers and users to promote and use both XBRL and Linked data, as each solution has a different role

to play. Combined use of them enables non-financial factors such as environmental and social performance of companies to be integrated into reasoning, allowing improved transparency and accountability by a diverse group of stakeholders.

Acknowledgments. This work was supported by the Systems Centre at the University of Bristol, the EPSRC funded Industrial Doctorate Centre in Systems (Grant EP/G037353/1) and the CDP Worldwide, London, UK.

References

1. Breslin, J., Passant, A., Decker, S.: The Social Semantic Web. Springer Science and Business Media, Heidelberg (2009)
2. Garcia, A., Gil, R.: Publishing XBRL as linked open data. In: World Wide Web Workshop: Linked Data on the Web (LDOW 2009), vol. 538 (2009)
3. Garcia, R., Gil, R.: Triplificating and linking XBRL financial data. In: Proceedings of the 6th International Conference on Semantic Systems. ACM (2010)
4. Goto, M., Hu, B., Naseer, A., Vandenbussche, P.: Linked data for financial reporting. In: Proceedings of the 4th International Conference on Consuming Linked Data, vol. 1034, pp. 123–135. CEUR-WS (2013)
5. Kämpgen, B., Weller, T., O’Riain, S., Weber, C., Harth, A.: Accepting the XBRL challenge with linked data for financial data integration. In: Presutti, V., d’Amato, C., Gandon, F., d’Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 595–610. Springer, Cham (2014). doi:[10.1007/978-3-319-07443-6_40](https://doi.org/10.1007/978-3-319-07443-6_40)
6. Madlberger, L., Thöni, A., Wetz, P., Schatten, A., Tjoa, A.: Ontology-based data integration for corporate sustainability information systems. In: Proceedings of International Conference on Information Integration and Web-Based Applications. ACM (2013)
7. Mora-Gonzalbez, J., Mora-Rodriguez, M.: XBRL and integrated reporting: the Spanish accounting association taxonomy approach. *Int. J. Dig. Acc. Res.* **12**, 59–91 (2012)
8. Namespaces in XML 1.1, 2nd (edn.) (2016). [goo.gl/UEMDG8](https://www.w3.org/TR/xml-names/). Accessed March