CAPSICUM Business Architects

Final Report

Prepared for:

Australian Taxation Office

Prepared by:

CAPSICUM Business Architects

E: troach@capsi.com.au

1st of June, 2015

Table of Contents

TABLE OF CONTENTS 2

INTRODUCTION 4

EXECUTIVE SUMMARY 5

TITLE: 5

PURPOSE OF PAPER 5

RECOMMENDATIONS 5

DETAILED RESPONSE AND JUSTIFICATIONS 8

SBR OBJECTIVES 8

SBR CURRENT STATE 8

SBR FUTURE DIRECTION 9

SBR SEMANTIC REPOSITORY (PROPOSED) 10

ENGAGEMENT SCOPE AND APPROACH 12

PREMISES 12

FINDINGS 13

RECOMMENDATIONS 14

RECOMMENDATION 1 14

INFORMATION ARCHITECTURE PRINCIPLES FOR A WHOLE OF ECONOMY DIGITAL INFORMATION EXCHANGE 14

PRINCIPLE 1: SEPARATION OF SEMANTICS AND SYNTAX 15

PRINCIPLE 2: SINGLE, CONSISTENT, REUSABLE DEFINITIONS 16

PRINCIPLE 3: BEYOND TAXONOMIES TO ONTOLOGIES 17

PRINCIPLE 4: FEDERATED ALIGNMENT OF VOCABULARIES 18

PRINCIPLE 5: A STANDARD IS ONLY AS GOOD AS ITS ADOPTION 19

RECOMMENDATION 2 19

DEPLOYMENT AND MAINTENANCE OF AN INFORMATION ARCHITECTURE USING A DISTRIBUTED/FEDERATED APPROACH 19

RECOMMENDATION 3 20

CONSTRUCTION OF DATA FOR DIGITAL TRANSMISSION 20

RECOMMENDATION 4 22

APPROPRIATE TOOL BASE TO USE 22

PROOF OF CONCEPT 25

PROTOTYPE 'REPORTING DICTIONARY' 25

POC FINDINGS 38

GOVERNANCE OF A SEMANTIC MODEL **39**

COMMON CORE DEFINITIONS SUPPORTING CROSS-DOMAIN REQUIREMENTS **39**
VERSIONING AND RELEASE STRATEGIES **40**

Introduction

The Australian Tax Office (ATO) under the following request for offer has engaged CAPSICUM BUSINESS ARCHITECTS PTY LTD.

Request for Offer:	BRR 15.5
	Consultant to provide recommendations in relation to Information Architecture principles for whole of economy digital information exchange

The request for offer was originally framed around the following questions;

Question 1	What are the Information Architecture principles for whole of economy digital information exchange?
Question 2	How do we deploy and maintain the Information Architecture using a distributed/federated approach?
Question 3	How should the data be constructed for digital transmission?
Question 4	What's the appropriate tool base to use?

Engagement period: from February 2015 to June 2015.

Executive Summary

Title: **Recommendations and Information Architecture principles for whole of economy digital information exchange**

Issue date:	8 June 2015		
Sponsor:	Andrew Joyce – Assistant Commissioner	**Business line/ branch**	BRR
Author:	Dr. Terry Roach – CAPSICUM Business Architects Pty Ltd	**Author phone:**	0421 054 804

Purpose of paper

Final paper from consultancy (Ref BRR 15.5) to provide expert advice in relation to the appropriate information architecture to support digital information exchange in Australia.

Recommendations

Recommendation 1: **Information Architecture principles for a whole of economy digital information exchange**

The expanding scope of the SBR programme to incorporate business domains beyond financial regulatory reporting, requires consideration of a broad set of interoperability concerns that influence the information architecture for the **content** of the SBR reporting dictionaries (the business vocabulary or *Semantics* of the standard) and the **protocol** for the messages exchanged (the grammar or *Syntax* of the standard).

Key principles that will protect the SBR assets and maximize their value and adoption by the reporting communities are:

Principle 1: ***Separation of Semantics and Syntax***

The fundamental priority of a reporting standard should always be to guarantee the semantic integrity of the business meaning in a message exchange. The syntax of the message is only a technical facility supporting accurate communication.

Locking the semantics into the syntax imposes on the community a specific language for the exchange, resulting in costly technical overheads for message translation at either end. Tight coupling of the standard to a particular language has disruptive implications whenever the business vocabulary evolves, with minor updates requiring expensive system upgrades for the entire ecosystem.

Principle 2: ***Single, Consistent, Reusable Definitions***

All vocabulary elements should be uniquely defined, consistently utilised complex elements should reuse simple elements. This means that the standard should be built on canonical element definitions i.e. atomic, fully decomposed element definitions that are expressed in their most simple form. The most canonical data structure is a triple (subject-predicate-object) e.g. Concept-attribute-Datatype or Concept-relatedTo-Concept.

Principle 3: ***Beyond Taxonomies to Ontologies***

A Taxonomy is a classification schema based on an iterative specialisation of subClasses. Taxonomies are extremely useful, very common and easy to comprehend, but have the limitations that they don't describe the relations between elements in the standard. Ontologies extend the modelling paradigm of a Taxonomy by expressing any type of relation between Classes. This aggregates enormous additional meaning and value into the model, extending it beyond a simple dictionary to a comprehensive semantic conceptualisation of the domain.

Principle 4: ***Federated Alignment of Vocabularies***

A federated alignment of information models allows the members of a domain community to participate in a collective exchange of information with a minimum of overhead. It facilitates the independence of community members to continue with their individual solutions, on different platforms and on independent lifecycles but still subscribe to common semantic definitions.

Principle 5: ***A Standard is only as good as it's Adoption***

Ease of adoption is directly related to the degree of effort and consequently cost that a standard imposes on it's adopter. Syntactic dependence on particular protocol has significant cost implications on the adopters. The ideal scenario is if the standard is syntactically agnostic and can be consumed and adopted in a variety of protocols.

Recommendation 2: Deployment and maintenance of an Information Architecture using a distributed/federated approach

A key enabler of a federated architecture lies in the abstraction of modelling layers which allows for multiple logical designs and schemas to built from a common conceptual model.

Maintaining the core business assets of the information vocabulary in the semantic layer, is a key to data federation information allowing for information to be exchanged and reports to be submitted in any number of formats (eg XML, JSON, XBRL) without any loss of semantic integrity.

Recommendation 3: Construction of data for digital transmission

As described above, when a common structure for meaning is kept at a higher layer of abstraction, variation in the protocols for transmission can be easily addressed at the syntactic layer in multiple formats and schemas (e.g., XBRL, JSON, XML). Since data in any digital format can be expressed unambiguously as triples (a *value* for a *property* of an *entity*), an RDF triple is a formalism for interpreting heterogeneous digital information. As a result, information in any schema can be posted and validated in any target format without the data providers having any particular understanding or even awareness of RDF as the protocol intermediary.

Recommendation 4: Appropriate tool base to use

Our strong recommendation for the technology foundation of a whole-of-economy information exchange is a semantic technology platform. The W3C Resource Description Framework (RDF) and Web Ontology Language (OWL) offer proven and accepted foundations for the universal identification, description, inference and querying of digital resources that is lightweight, flexible and extraordinarily expressive.

At the syntactic level, RDF is a very simple, lightweight, open protocol that is increasingly being adopted for the exchange and sharing of semantic information. This will allow the SBR standards to be readily integrated into the growing community of linked open data. The OWL ontology language provides a powerful formal logic system for defining complex business rules and restrictions.

Detailed Response and Justifications

SBR Objectives

The SBR program was created to reduce regulatory burden and simplify government reporting processes by facilitating the automation of reporting obligations from business to government, through the use of open standards and harmonising reporting terms.

SBR Current State

The SBR architecture is closely coupled with the current SBR technologies, a central component of which is XBRL (eXtensible Business Reporting Language). The binding of the SBR architecture to a specific syntax is seen as a potential obstacle to increasing uptake of SBR beyond the current user base.

Findings from workshops held by the Taxonomy Extension Working Group (Sept/Oct 2014) included the following, which acknowledge the challenge to adoption resulting from the exclusive usage of XBRL;

Findings (direction of SBR taxonomy extension)[1]

1. XBRL is a complex language. It is not well suited beyond the reporting of financial regulatory information, but for non-financial reporting, there are international standards in use by other agencies that are not XBRL-based. There is no practical benefit to re-implement their taxonomies in XBRL.
2. XBRL is also not well suited for mobile platform and bulk data lodgement.

[1] SBR Board submission Oct. 2014 (Agenda Item 9.7: Attachment F)

SBR Future Direction

To summarise the preceding section, the deficiencies of the existing SBR taxonomy relate to complexity and inflexibility in its technological implementation and narrowness of scope.

- Simplification

  It is recognised that the cost of development and maintenance of the SBR taxonomy is high, largely due to its high complexity.

- Adoption and scope

  Barriers to adoption of SBR by software developers and by reporting businesses exist, and the level of adoption is low.

  One acknowledged barrier to adoption is the prescribed use of the XBRL as the exclusively-supported reporting language.

  At present, the limited reporting scope of regulatory financial data is the key driver for the core principles supporting of the SBR architecture. The broadening of the scope of SBR beyond the financial regulatory domain is constrained by the current commitment to XBRL, the primary application of which is financial regulatory reporting.

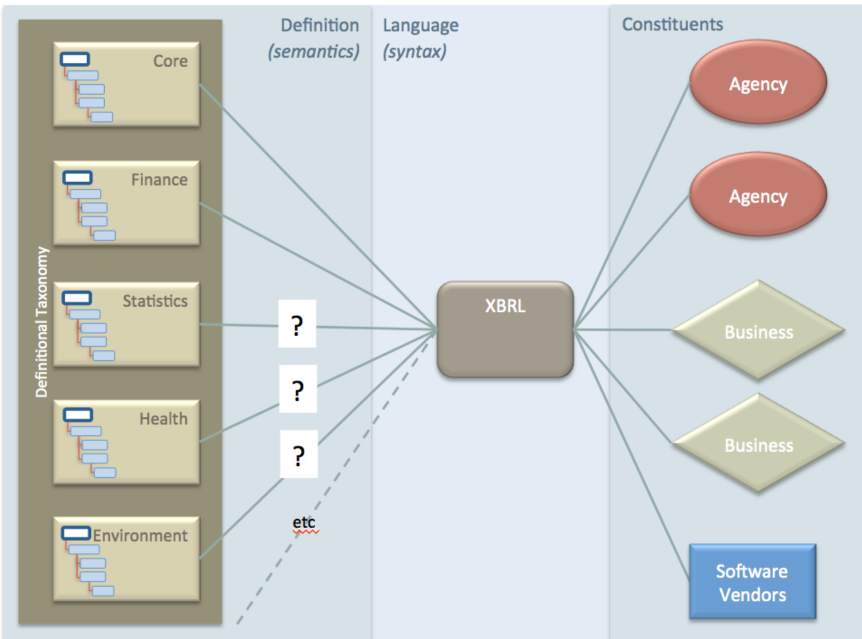
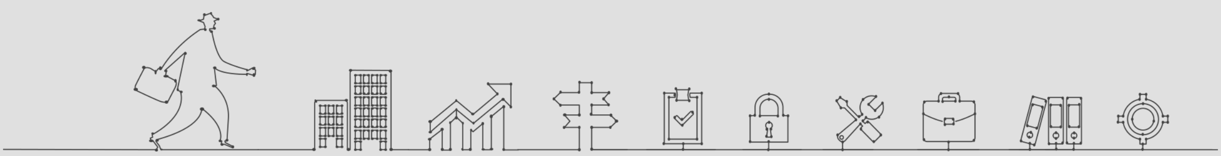


Fig. 2.1 SBR current state - XBRL is the prescribed reporting language across reporting domains



- Minimisation of reporting burden placed on businesses

  Currently under SBR, business-to-government (B2G) reporting takes place as a discrete event, separate to routine business activity. B2G reporting represents, therefore, a non-value-adding, compliance overhead to businesses.

  Part of the vision of the SBR Program is to enable businesses to fulfil their reporting obligations as a natural consequence of 'normal' business activity - such that routine reporting (ideally) occurs without incurring an associated overhead. Likewise, reporting should be of natural business data, and should not require the derivation of report-specific data by a business.

Summarised below are the features of the future vision for information management and validation within the SBR architecture;

- The capture of the semantics of business process relevant (directly and indirectly related) data
- Captured data is modelled semantically without the constraint of a particular syntax
- Semantic description of both information and related validation rules enabling improved quality/consistency in the application of rules
- Business reference models enable the alignment of captured information
- Information exchange can take place in the context of a business process, or can be document-based

Key themes taken from the above points:

- Separation of information content (semantics) from information structure (syntax)
- The need for greater completeness (or richness) of the information captured to provide a holistic view of business processes

SBR Semantic Repository (Proposed)

A key enabler for the success of SBR is the ability to maintain all of the necessary information required to complete a business process in a semantic layer that is kept logically distinct from the message exchange protocols. By representing this information at these different levels of abstraction, people with varying levels of technical and business knowledge can find common ground on which to interact.

This mechanism also supports syntactic independence and facilitates message exchange in any variety of technical schemas, enabling the generation of specific deployment representations of the information to suit the technology choices of particular communities without impacting on the technology choices of others. Regardless of the deployment choice, the information is always linked back to the semantic model.

Many communities will already have an existing preference for pre-established industry protocols for information exchange. For example, the Financial Regulatory Reporting community may continue to use XBRL as a message format, while the Health community can use HL7. Both domains will be able to subscribe to a mix of core, cross-industry semantic definitions as well as their own domain specific vocabulary. Cross-domain adoption is supported since these semantic definitions can be serialised in the protocol chosen for the particular domain.

SBR will be in a strong position to take advice from domain representatives and support any well-founded choices of protocol as well as to control against the proliferation of multiple protocols into a general free-for-all.

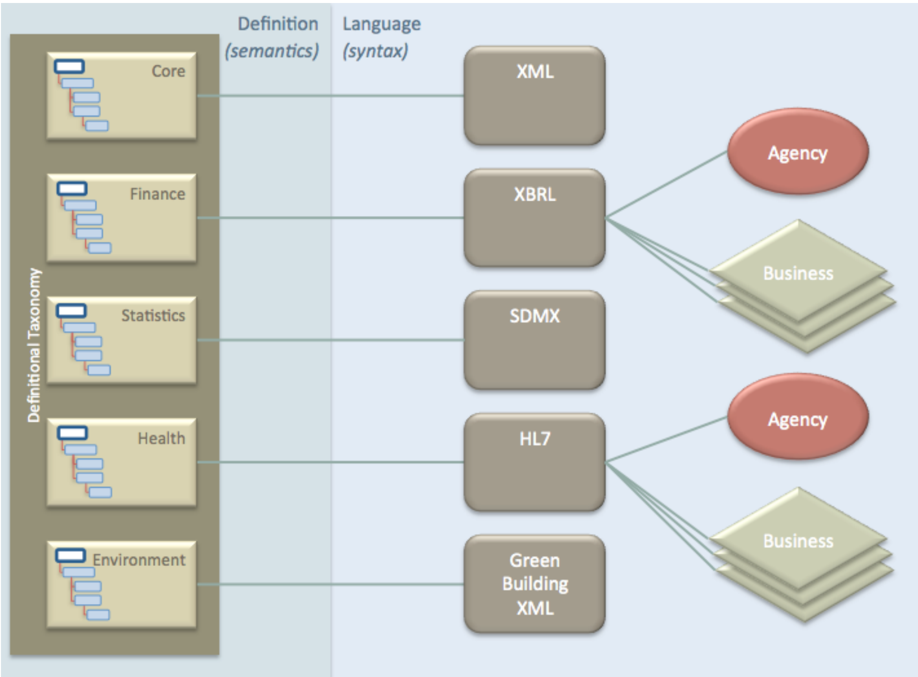


Fig. 2.2 SBR future state – Separation of information content (semantics) permits the use of reporting languages aligned to communities' technology choices

The information in the semantic repository can be consumed in a variety of ways depending on the audience. A software developer may prefer to download an information bundle into their software development environment, while a business user may be better off following a carefully designed learning package.

The semantic model will also express the business and technical requirements of offering and consuming digital services, and provide linkages to service providers who can help facilitate the process.

Engagement Scope and Approach

The initial scope of content was agreed as being that of 'employer reporting' with an initial focus being placed on the Payroll domain. To demonstrate the stated capabilities, content was provided by the participating agencies (ATO - SBR, ATO - Single Touch Payroll initiative, DHS) and this content used as the basis for initial modelling of the domain.

Employer related reporting lends itself to the SBR solution, including its machine to machine online gateway and harmonised reporting terms (taxonomy), as it enables the preparation and lodgement of multiple reports, often containing duplicate or similar information, to disparate agencies.

The difficulties for business in completing employment-related reporting requirements are often compounded by agency-specific reporting cycles, which are frequently not aligned with natural business processes. Most of the employer data reported to Government is derived from payroll systems.

Currently, B2G reporting imposes significant compliance costs where employing businesses (40 per cent of businesses) are required to report on a number of different employment-related matters (either general or industry-specific) to a range of agencies in Commonwealth or State jurisdictions

Premises

In addition to providing recommendations in response to the specific questions posed by the Request for Offer, this work seeks to demonstrate the following:

- A framework supporting the interoperability of co-existing domain-specific reference data standards (e.g. HL7 for health, UBL for invoicing, SDMX for statistical data exchange)

- The creation and maintenance of a single, common, language-independent definition of terms following semantic modelling principles
- The development of common definitions in both human readable and machine actionable forms
- The feasibility to generate specifications and payload formats in different languages (potential candidates to be determined) to ensure data exchange utilises the common core terms and definitions

Findings

Findings are organised under the following sections;

Recommendations

- Recommendations formulated in response to the questions posed in the formal Request for Offer

Proof of Concepts (PoCs)

- Description and evaluation of the PoCs undertaken to supplement the responses to Request for Offer, and to demonstrate the additional premises put forward.

Recommendations

Recommendation 1 Information Architecture principles for a whole of economy digital information exchange

As the SBR programme expands beyond the Financial Services (regulatory reporting) domain to address information exchange in other industries, new architectural issues surface. A whole-of-economy approach requires consideration of a much greater set of concerns to arrive at an information architecture design that will elegantly facilitate this broader context, addressing greater diversity of semantic and syntactic requirements without a corresponding increase in complexity.

The exchange of digital information, by its nature, requires several points of agreement between the parties relating to how the information being exchanged will be structured, transported and mutually understood. Issues relating to the actual transport mechanisms and the "logistics" of the exchange are beyond the scope of this consultancy. The focus of this work is on principles relating to the structure and meaning of the information exchanged.

The challenge is one of information interoperability and it is a common one. Many industry domains operate as common ecosystems where sharing information between entities is fundamental to the industry endeavour. Financial markets for example are an entirely digital industry predicated on the exchange of financial instruments and payments by digital means. Similarly the health community has many touch-points where patient information is exchanged between health providers and there are similar requirements for information exchange in many other industry domains within the SBR scope.

Participants in such communities have been tackling these interoperability challenges for a long time and the common approach to addressing them is for the industry to agree on information standards or reference models. It is quite common for several competing or overlapping standards to develop and for a domain to have a fairly fragmented set of information standards of varying maturity and in various states of adoption. They all have one thing in common; a proposal for a common vocabulary and grammar for the domain in a way that will be acceptable and useful to the entire community. The problem is that each attempt at standardisation runs the risk of being self-defeating in that it generally results in yet another conflicting point-of-view that needs to be reconciled.

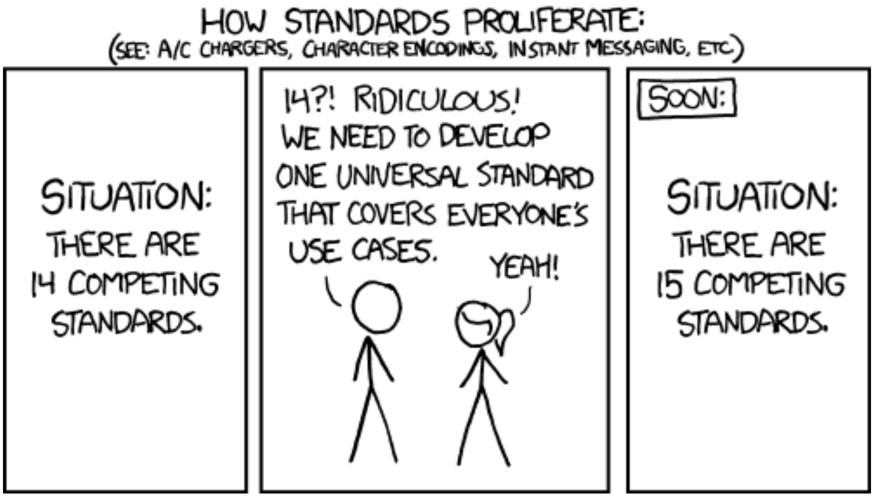


Fig. 3.1 How Standards Proliferate

Source: https://xkcd.com/927/

Faced with these concerns and given this diverse, fragmented landscape, what are the over-riding architectural principles that should guide a common whole-of-economy framework for information exchange?

Principle 1: Separation of Semantics and Syntax

The key purpose of an information standard is to enable interoperability in the sharing and exchange of data by establishing:

- agreement on the ***content*** and
- specifications for the ***structure*** of

a controlled vocabulary.

- The ***content*** of an exchanged message is known as the message **Semantics** and expresses the meaning in the communication. Semantic definitions are defined by business subject matter experts and establish a common vocabulary for the domain.
- The ***structure*** of a message is expressed in the message **Syntax**, i.e. the format, grammar or protocol of the language that is used to encode the message for exchange. Syntactical definitions are typically chosen by technical teams and take into consideration, technical constraints and requirements of the exchange environment.

It's clear that the Semantics and Syntax of a standard are quite separate subjects and very important that each of these considerations be managed and treated independently of each other.

The fundamental priority of a standard should always be to guarantee semantic integrity in the meaning of a message. The syntax is only the messenger.

Technical requirements for syntax that may seem very appropriate at a particular point in time can often prove quite fickle and volatile, subject to constantly shifting technology fashion and ongoing technology innovation.

Publishing a vocabulary that is locked into a particular protocol inevitably means that the community participants are entirely committed to both the meaning and the format of the exchanged message, and accept the translation overheads of message conversions and message mappings that may eventually be required at either end of an exchange.

Locking the semantics into the syntax can also have disruptive implications when the semantics of the standard evolve, as they inevitably do. If the standard is tightly coupled into the language, minor updates to a vocabulary definition can require the entire ecosystem to have to conduct expensive system upgrades each time the vocabulary evolves.

Principle 2: Single, Consistent, Reusable Definitions

It is an absolute and fundamental requirement of an information standard that any vocabulary element is uniquely defined, consistently utilised and that complex elements reuse simple elements.

An important strategy that will assist with this is to construct the vocabulary from canonical element definitions (i.e. atomic, fully decomposed element definitions that are expressed in their most simple form). Decomposing element definitions to a canonical form will provide the greatest flexibility in constructing a standard. A standard that can be reduced to fully decomposed elements facilitates usage and adoption by the community members, by the greater flexibility it affords in mapping the standard to more complex definitions in a member's own information architecture.

In the case where a set of standards will potentially cross domains, this requirement becomes even more significant. A simple, typical example is an Address definition. Dozens of approaches exist for modelling Addresses, ranging from quite unstructured (e.g. Address Line 1) to fully decomposed (AS4590) influenced by any number of regional, domain specific, technical or business considerations, for example. It is of course possible to map a highly decomposed Address definition to a system that uses unstructured Addresses, but would be quite impossible to do the opposite.

The most canonical data structure of any of the traditional data-modelling paradigms is a triple (subject-predicate-object e.g. Concept-hasAttribute-Datatype or Concept-relatedTo-Concept). Triples form the basis of the W3C Semantic Technologies stack built on the Resource Description Framework (RDF).

Maintaining the vocabulary semantics in RDF triples will go a long way to ensuring that the specification is as decomposed as it could possibly be and forms a good basis for subsequent translation into any other syntax required.

The triple construct gives us a clear and simple set of foundation constructs for the structural composition of an definitional element in an information standard, namely:

- **Classes** (interchangeable with Concepts, Entities or Objects depending on the modelling paradigm)
- **Properties** (of two major types, Attributes or Relations)
- **Datatypes** (reusable definitions that constrain the possible values for a Property)

Principle 3: Beyond Taxonomies to Ontologies

Many information standards are constructed in the form of hierarchical Taxonomies. A Taxonomy is a classification schema based on an iterative specialisation of Class definitions, where Classes represent sets of things in a domain that have common properties. Specialisation is achieved by subclassing a Class to further distinguish the members of the Class based on finer grained distinctions of the member properties. A defining feature of a Taxonomy is that the only possible relationship between Classes is a subclass relation, representing continuous specialisation.

Taxonomies are extremely useful, very common and easy to comprehend. A common example is a classification of species. The limitations of a Taxonomy are that they don’t provide information on any other possible relations between elements in the model (other than subclasses) and are restricted to explaining elements that are all necessarily members of a common root Class.

Ontologies extend the modelling paradigm of a Taxonomy by allowing the definition of any other types of relation between Classes. Ontological relations provide a much greater depth of expression, contextualising the model by explaining how elements interact and participate in the domain through associations of items from within or even across different domain models.

Ontological relationships can be traversed to draw inferences about existing knowledge, thereby creating new knowledge. For example if two repositories contain different information about the same item, this knowledge can be combined or otherwise used to establish additional facts.

By expressing an information standard as an Ontology rather than a Taxonomy, enormous additional meaning and value is aggregated into the standard, enabling the standard to extend beyond a simple vocabulary to a comprehensive semantic conceptualisation of a domain.

Principle 4: Federated Alignment of Vocabularies

A standard is only required in situations where the proliferation of distinct vocabulary definitions is a risk or is already the case. Through the natural evolution their own colloquial vocabulary or through adopting different information systems or business methods the members of a domain will inevitably have evolved their own heterogeneous business definitions and technical specifications for expressing the meaning and organising the structure of their information. The purpose of a standard is not to impose neither a vocabulary or a language on the community, but to allow the community to align their vocabularies and interpret messages from other vocabularies within the domain.

A federated alignment of information models allows the members of a domain community to autonomously coexist using their current information vocabularies and systems but participate in a collective exchange of information with a minimum of overhead.

An information standard is never intended to provide an optimised data-model for operational purposes. It is generally the case that there are many unique, possibly conflicting considerations that individual participants in a domain may need to be concerned with in the design of their information systems which would make it highly improbable that a single data-model design would be appropriate for all of the possible usages of a domain vocabulary. The diverse business problems that community constituents face will invariably require unique specialised information model designs.

The principle benefit of a federated approach is that it facilitates the independence of community members to continue with their individual solutions, on different platforms and on independent lifecycles but still subscribe to common semantic definitions.

Principle 5: A Standard is only as good as its Adoption

Despite the best intentions, an information standard only has value when it is adopted by the domain community. There are several well known examples of information standards that have been widely acclaimed for the elegance and purity of their design but have failed in their adoption by the community.

For example, the health standard HL7 RIM v 3.0 is highly regarded in the health standards community but despite being widely promoted for nearly 20 years, adoption has been low due to it's very abstract nature. In contrast HL7 FHIR, in only it's second year of publication has already gained significant traction as a pragmatic and useful standard that is well aligned to industry requirements.

Of course there are many factors that will influence adoption, but one of the most obvious is the ease with which constituents can embrace and makes use of the standard. Ease of adoption is directly related to the degree of effort and consequently cost that a standard imposes on it's adopter, as well as the ongoing effort in maintaining the adoption.

Syntactic dependence on particular protocol is clearly a significant barrier since it implies either costly conversions in re-architecting solutions and/or costly real-time translations of messages as they are exchanged. The ideal scenario for an adopter is if the standard is syntactically agnostic and can be consumed and adopted in a variety of protocols.

Recommendation 2	Deployment and maintenance of an Information Architecture using a distributed/federated approach

The value in a federated information model is that it provides a high degree of autonomy for the constituent systems but strong alignment to a common reference model. Effectively that implies that multiple heterogeneous implementations can subscribe to a common conceptual design, which is precisely the goal of a whole-of-economy Information Standard.

A key enabler of a federated architecture lies in the abstraction of modelling layers which allows for multiple logical designs and schemas to built from a common conceptual model. Classic data-modelling theory prescribes 3 well accepted modelling layers (Conceptual, Logical and Physical) as does Model-Driven-Architecture

(Computationally Independent Model – CIM, Platform Independent Model – PIM, Platform Specific-Model PSM). In the context of the discussion in Principle 1 above, these traditional modellng layers can be loosely described as a Semantic Layer, Logical Layer and Syntactic Layer.

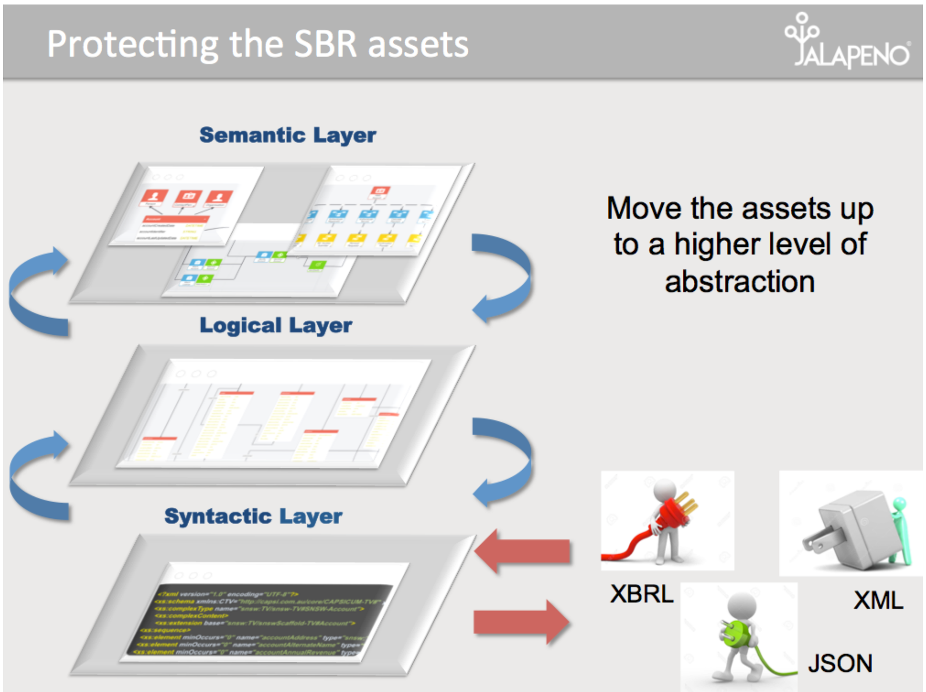


Fig. 3.2 Protecting the SBR Assets

Maintaining the core assets of the information standard in the Semantic layer, is a key to data federation, allowing for multiple different normalisation strategies for the ERD tables at the logical layer and the generation of schemas in any number of formats and from any number of logical designs without any loss of semantic integrity.

Using this approach, information to be exchanged and reports to be submitted in a number of formats (eg XML, JSON, XBRL), which allows for the SBR vocabularies to be serialised in the industry protocol that is appropriate for a particular domain.

Recommendation 3 Construction of data for digital transmission

Data transmission in a whole-economy distributed system has to navigate between two conflicting requirements. On the one hand, we have to allow for differing innovation schedules for various agencies and regional governments. It is unreasonable to expect every entity that has to report data to upgrade their systems simultaneously. Therefore, heterogeneity in reporting systems will be the norm, not the exception.

On the other hand, all of these systems have to be able to interoperate in terms of content. We can't have different reporting requirements for different regions just because they have not yet migrated to a new technology. It must be possible to do comparative analyses across jurisdictional boundaries.

This impasse manifests in two ways – one having to do with the format of the data that is being shared, the other with its content.

The solution to this impasse in terms of format is based on metadata management, as outlined above, through the separation of the semantic and logical layers from the syntactic layer. A common structure for meaning is kept in the higher layers, while variation in presentation form (e.g., XBRL, JSON, XML) is achieved at the syntactic layer.

This separation applies to the data as well. Data that is constructed in any digital format (ranging from modern formats like JSON, XML and XBRL, to legacy formats like tab- or comma-delimited files) can be expressed unambiguously as triples - a *value* for a *property* of an *entity*. The metadata platform pictured in Fig 3.2 allows schemas for these formats (and others) to be coordinated at a semantic level. Each of these schemas can be used to post and validate messages in any of the target formats. It is important to note that while RDF, as a standard for managing data triples, provides a common framework for understanding and aligning data, that the providers of data must not be required to have any particular understand or even awareness of RDF as a technology. RDF is used as a formalism for unambiguously interpreting digital information from multiple sources, not as a digital medium itself.

The Capsicum framework also provides relief for the impasse in terms of alignment of data content. A common method for aligning content in many industries is the adoption of some standardized terminology. Such systems are commonly referred to as controlled vocabularies, code lists, data points or (as in the Capsicum framework) Value Sets. Many of them are highly standardized (e.g., the UNSPSC product codes, SIC industry codes, ICD diagnostic and clinical codes, the Westlaw key numbering codes, ISO country, language and currency codes, etc.), while others are more specific and less controlled (preferred customer levels, lists of media genres, age groups, etc.). Just like data formats, different reporting entities will have differing needs in terms of the vocabularies they use for these features.

The Capsicum framework provides a means for managing these Value Sets, relating the specific terms to one another and managing their identities and formats. Again, as shown in Fig 3.2, this management of content is independent of any particular syntactic presentation. Two digital messages can refer to the same code in a value set, while expressing that message in different formats.

The overriding principles for digital transmission of data in a distributed, whole-economy setting are the same as the principles for managing metadata – they are based on the foundation of separation of semantics from syntax, of medium from message. In the case of data transmission, this isn't just a desirable state of affairs – it is necessary in a distributed data sharing setting.

Recommendation 4 Appropriate tool base to use

As indicated throughout the above discussion of Principles and Recommendations, our strong recommendation for the technology foundation of a whole-of-economy information exchange is a semantic technology platform.

In recent years the efforts of the W3C consortium in promoting the Semantic Web and Linked Open Data initiatives have gained broad support and generated global momentum behind their standards for structuring semantic content in open, machine-readable and machine intelligible languages. Their Resource Description Framework (RDF) and Web Ontology Language (OWL) offer proven and accepted foundations for the universal identification, description, inference and querying of digital resources that is lightweight, flexible and extraordinarily expressive.

At the semantic level, converting the existing Taxonomies into Ontologies will be a relatively easy, one-off conversion exercise. The existing XBRL-based assets will be converted into RDF (as demonstrated in our prototype) and new semantic relationships will be established between the definitional elements in the Taxonomies. This will accrue significant additional value to the standard, converting it from a dictionary into a true conceptual model of the domain.

Ontological graphs are easily rendered as visual graphs, which makes the content of the model much more accessible to non-technical users. An RDF triple is also a very good construct for mapping Classes, Properties and Datatypes, which facilitates easy mapping of the SBR standards to other industry standards (e.g. HL7 for Health) that will greatly assist with SBR adoption in those industries.

The OWL ontology language provides a powerful formal logic system for defining complex business rules and restrictions.

Consumers of the standards (e.g. reporting entities) will also appreciate the facility of mapping the SBR concepts to their own conceptual models, which will be another factor that assists with adoption.

At the syntactic level, RDF is a very simple, lightweight, open protocol that is increasingly being adopted for the exchange and sharing of semantic information. This will allow the SBR standards to be readily integrated into the growing community of linked open data.

Conversion of other protocols to and from RDF is relatively easy, as we have demonstrated in our prototype, with RDF proving a very reliable and practical foundation for maintaining the standard but also for assisting business constituents to adopt and make use of the standard.

The Capsicum Framework is an example of a semantic meta-model that offers significant potential for extending the SBR scope beyond just a reporting information standard to demonstrate how reporting obligations can be mapped to business process and governed by business rules. The current desire to move towads continuous reporting cycles (eg the ATO Single-Touch Payroll initiative is a great example of this)

In summary, to achieve the objectives outlined in the Recommendations above, the proposed toolset should address the following minimum capabilities:

- Ontology creation and maintenance
- Version control and release management
- Separation of modelling layers, (at a minimum syntactic from semantic layers)
- Conversion of models between modelling layers (Model-Driven-Engineering)
- Graphical modelling notations and visualisation
- Ability to search/browse/query the models
- Ability to generate custom reports
- Ability to import and export model content in multiple formats (CSV, PDF, XML, JSON & XBRL at a minimum)

- Ability to provide custom views for a variety of audiences including both business and technical stakeholders
- Ability to express business rules and business processes
- Ability to manage the curation of codelists (ValueSets)
- Ability to map Concepts, Properties and Datatypes to other information and message standards

Proof of Concept

A Proof-of-Concept was undertaken to demonstrate the capabilities and potential opportunities for value that are offered by a Semantic Technology Platform. The PoC made use of the CAPSICUM Framework, a semantic meta-model for business architecture and the Jalapeno semantic modelling platform.

Two streams of work were followed:

1. Development of a prototype 'Reporting Dictionary' allowing the discovery, visualisation and output of definitional and reporting taxonomy elements.

   The reference for the prototype was the existing Australian Reporting Dictionary (http://dictionary.sbr.gov.au).

2. Modelling of content in the domain of Employer Reporting including alignment of an Reporting reference model with an industry standard for Human Resource Management (HR-XML).

Prototype 'Reporting Dictionary'

The existing Australian Reporting Dictionary is a browser-based tool through which a user can search for, and view details of elements within the SBR definitional and reporting taxonomies.

A subset of SBR taxonomy content was used to develop the prototype; the Individual Income Tax Return (IITR) and its related definitional elements.

The 'Reporting Dictionary' prototype, demonstrates the following capabilities;

- The ability to provide a means to query and present SBR taxonomy content in a form which is understandable to a broad audience.
- The ability to render taxonomy item definitions in specific output formats (e.g. XSD, XBRL, RDF)

Landing Page

The landing page of the prototype was constructed to show a search bar, descriptive introductory text and a graphic providing further explanation of the taxonomy components and their respective terminology.

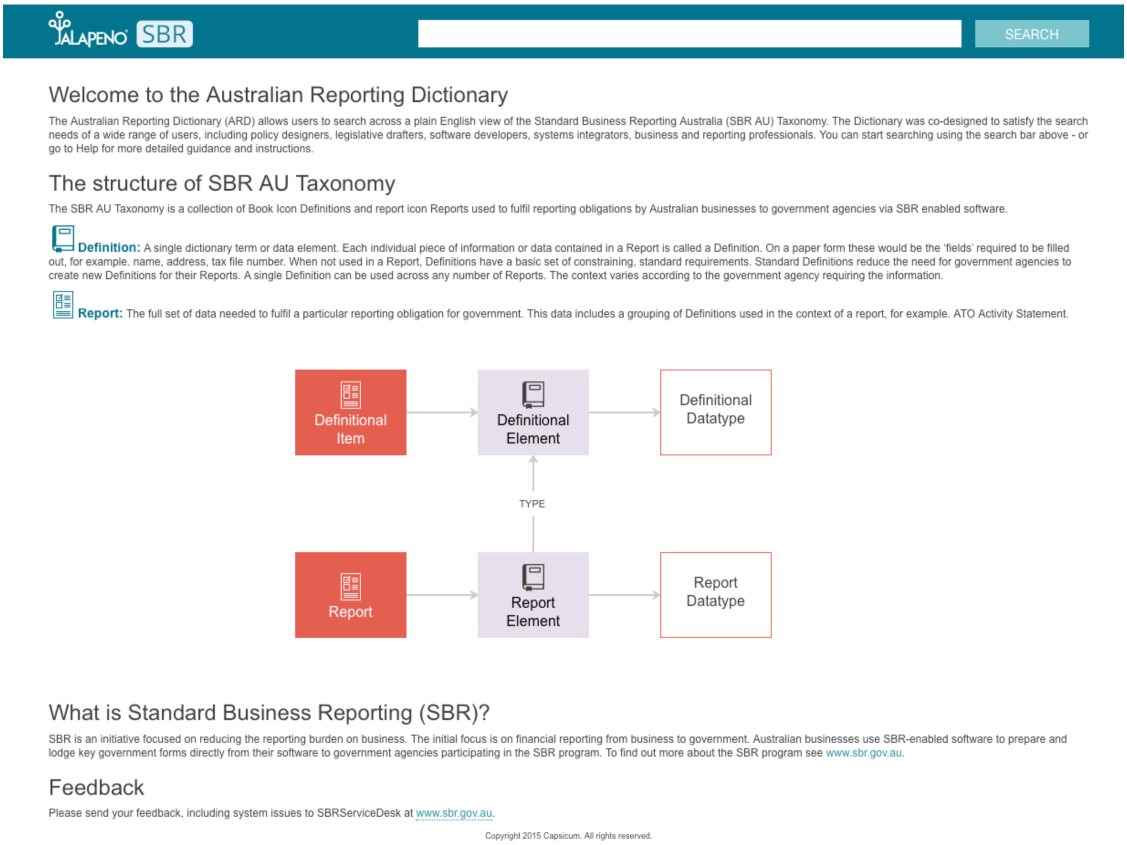
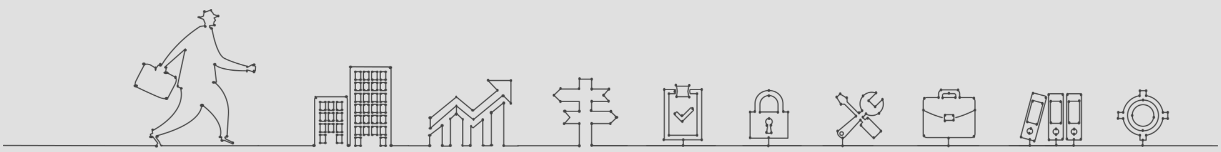


Fig. 3.3 Prototype 'landing page'

Search

On performing a search, the matching terms (along with their URI's) are listed for selection (see the example shown in Fig 3.4). A term may be a definitional item, definitional element, report or a report element.



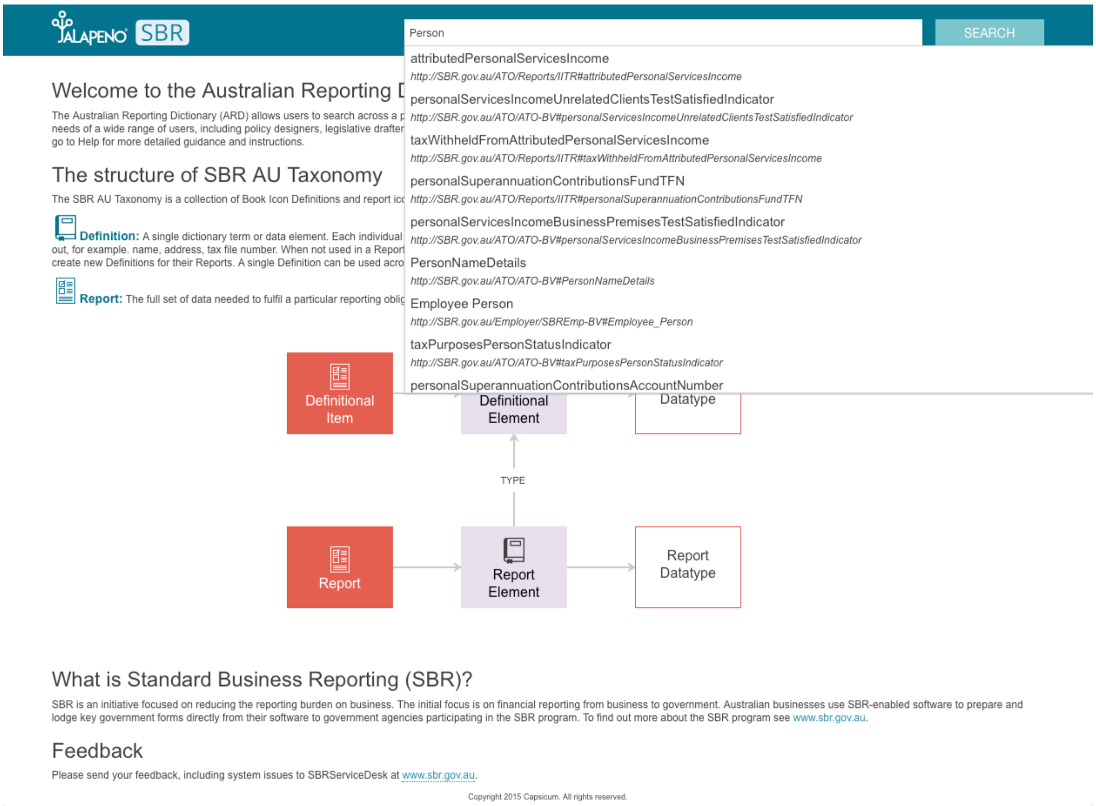


Fig. 3.4 Text-based search

Summary Tab

Once a term is selected from the search results, details of that term are presented on the 'Summary' tab. The specific content of the Summary is dependent upon the type of term selected (compare the example definitional element summary shown in Fig 3.11 and the example report element summary shown in Fig 3.10).

The position of the term in the taxonomy is shown via a 'browse-able' tree structure. Properties of the term are also given.

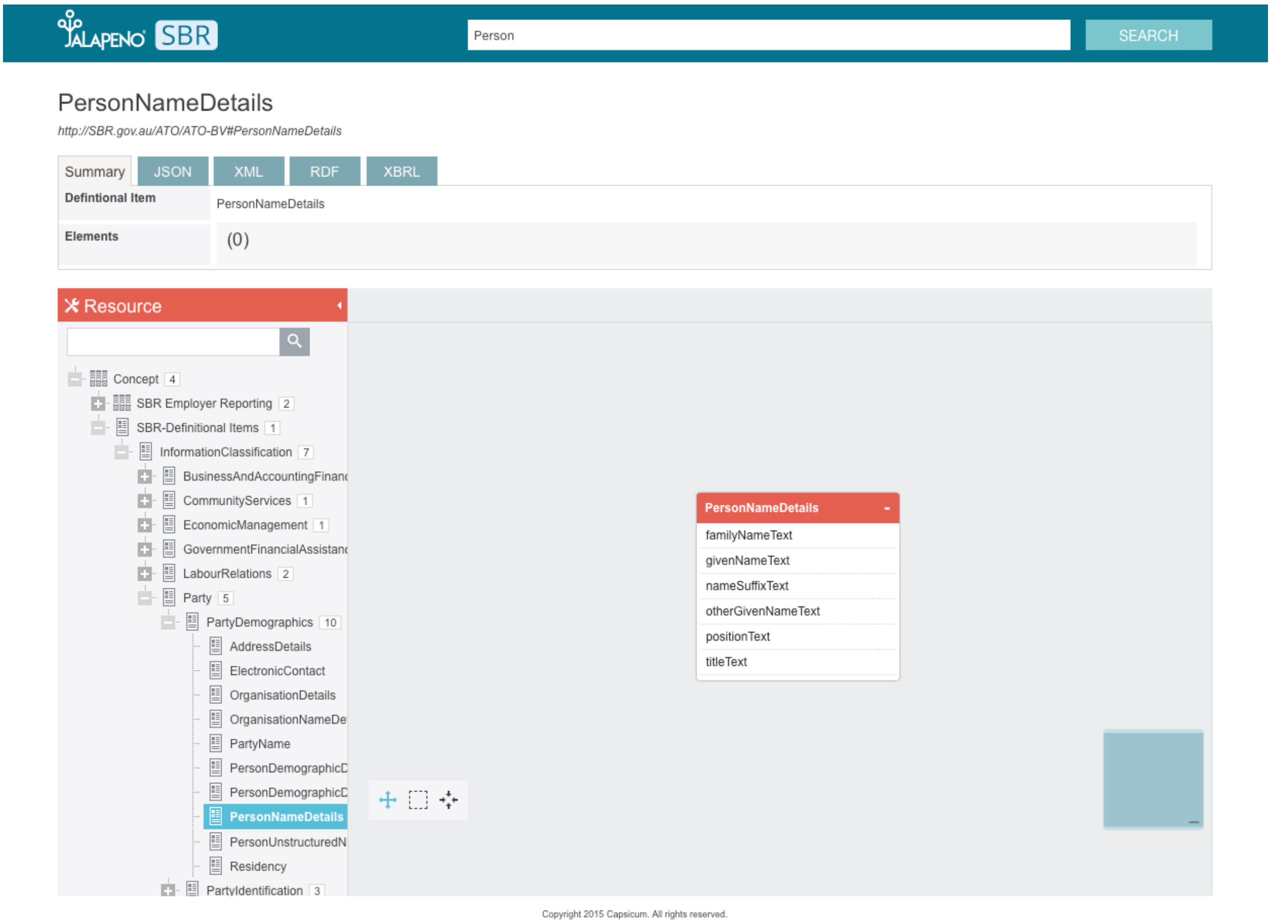


Fig. 3.5 Graph display of search result (definitional item example)

Data Formats (syntax)

The tabs JSON, XML, RDF, XBRL present the term as represented in the data format as indicated by the tab name. See Figs 3.6 – 3.9.

The tabs demonstrate the conversion and rendering of the underlying semantic model in multiple formats according to the requirements of the user.

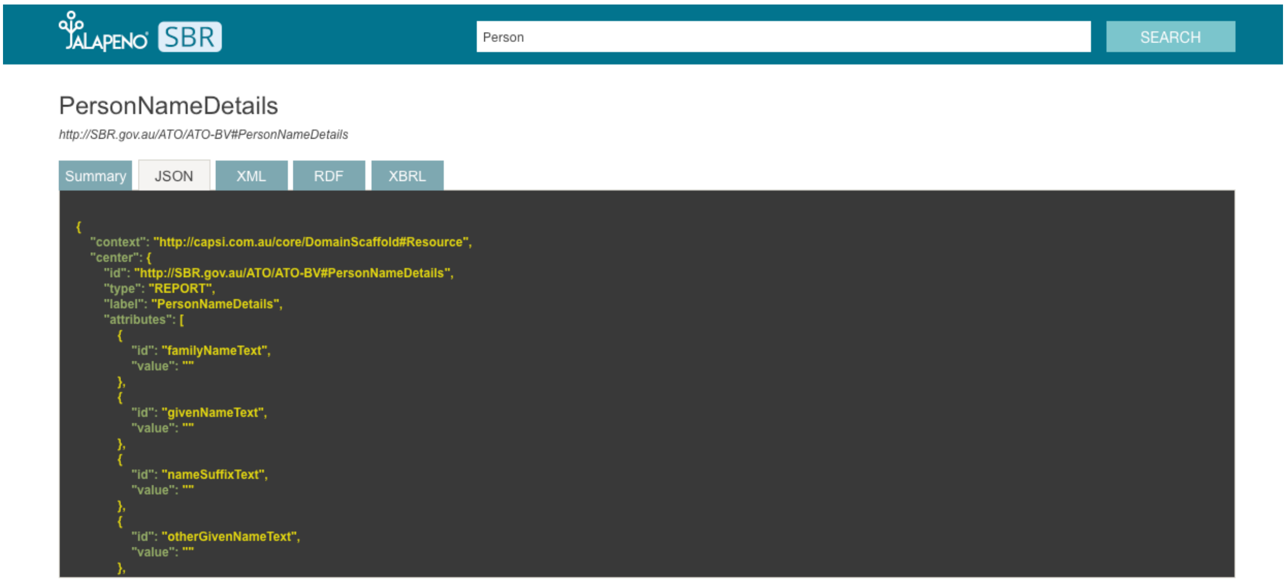


Fig. 3.6 Search result rendered as JSON

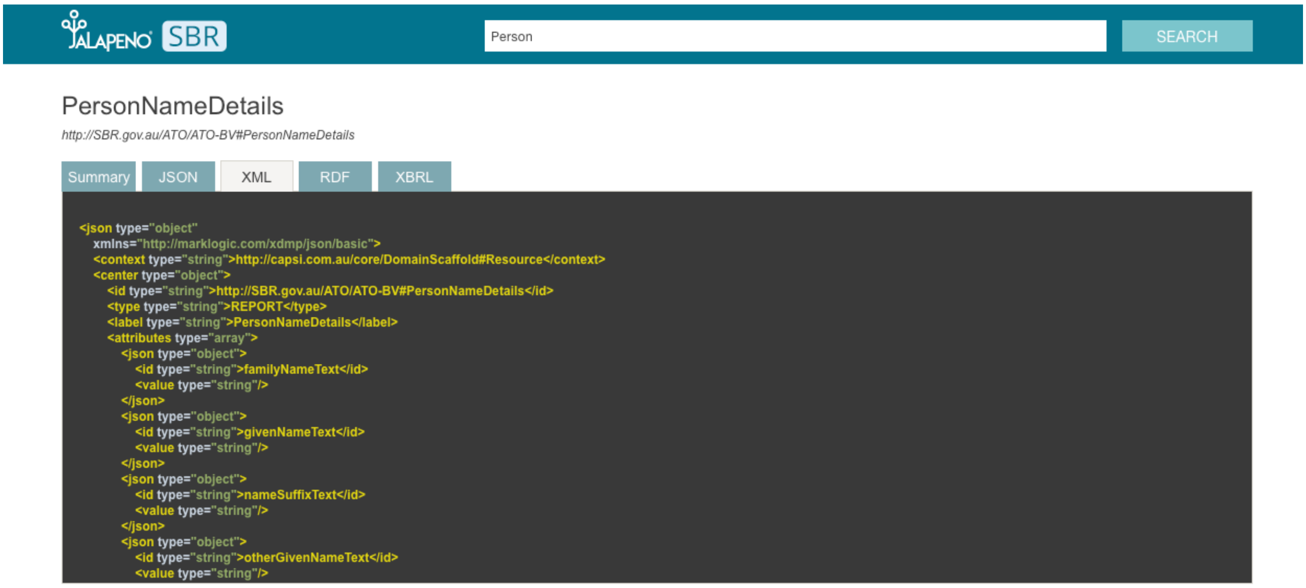
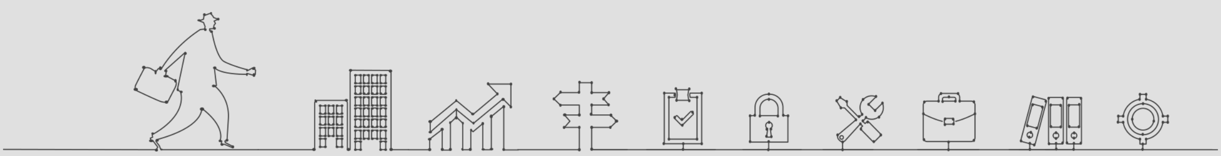


Fig. 3.7 Search result rendered as XML



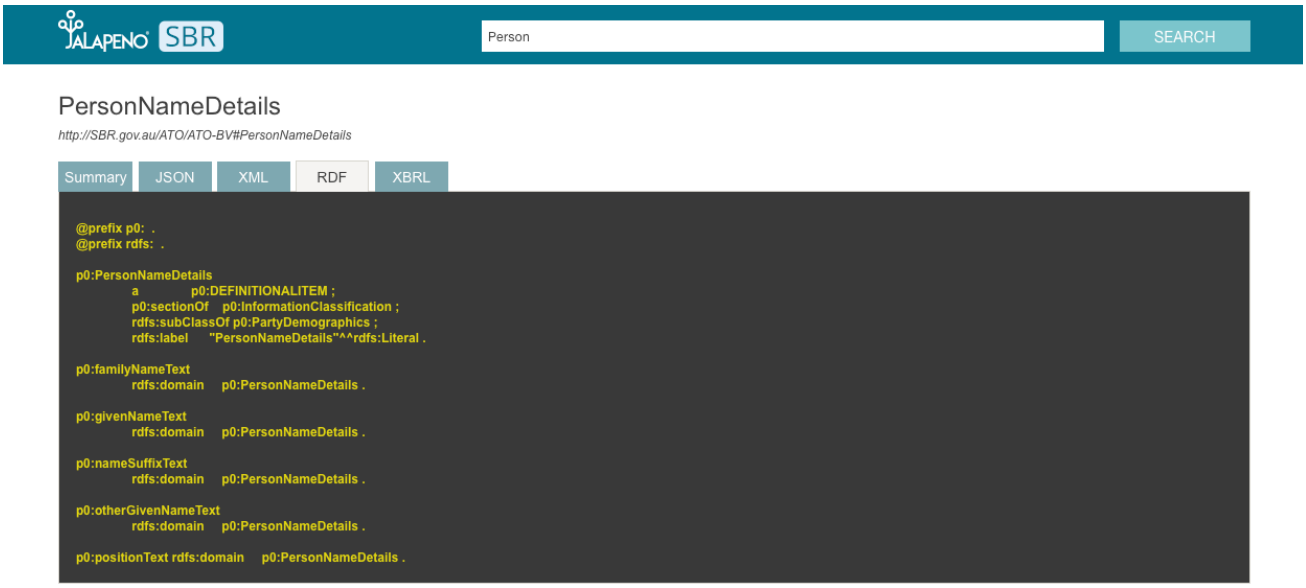


Fig. 3.8 Search result rendered as RDF

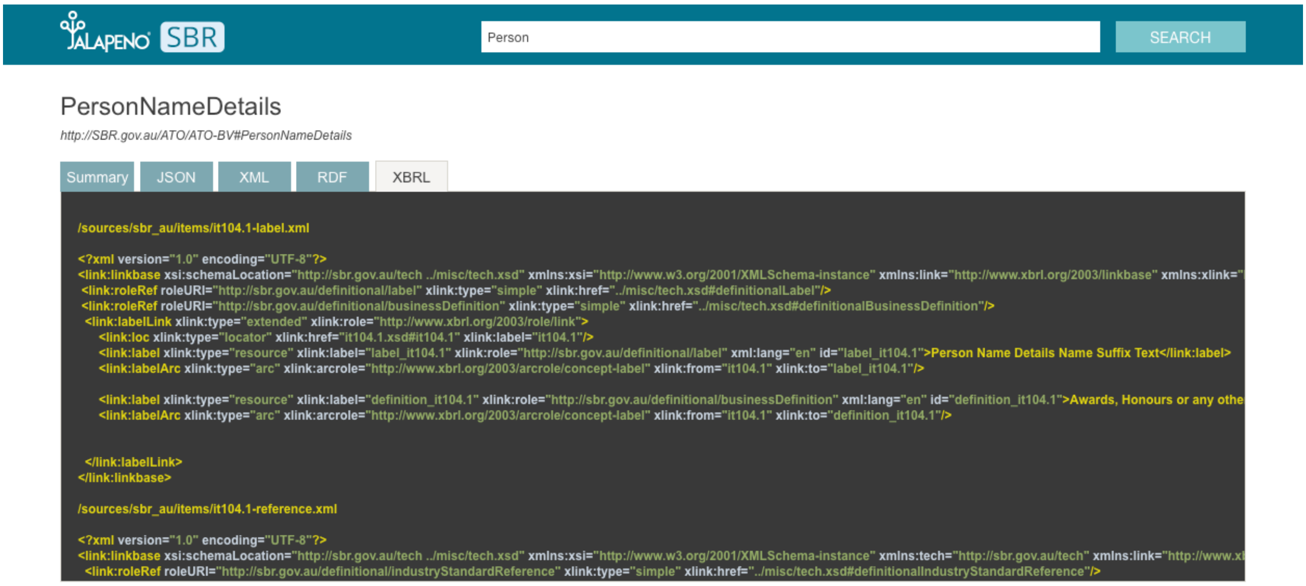
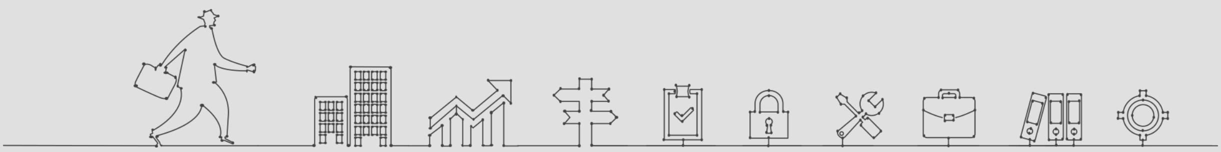
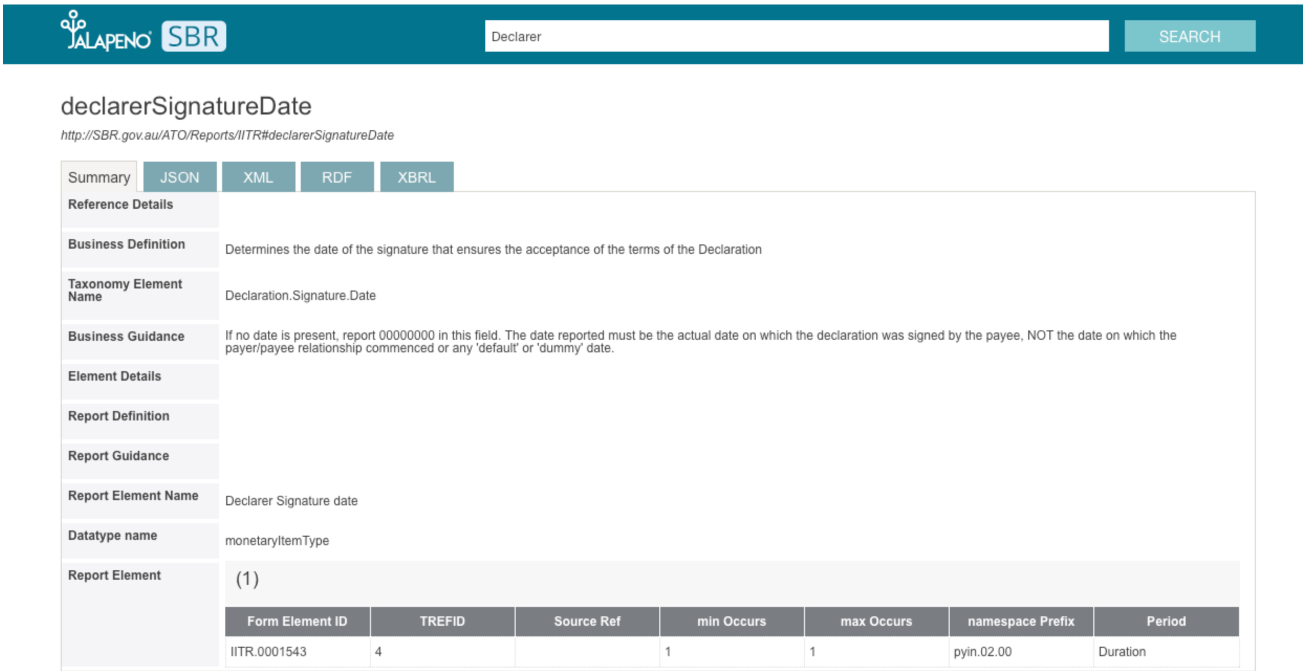


Fig. 3.9 Search result rendered as XBRL

Report Element Summary

For a report element, the Summary tab contains a profile of the element including guidance information, and references to the related definitional element (see Fig 3.10).





JALAPENO SBR | Declarer | SEARCH

declarerSignatureDate

http://SBR.gov.au/ATO/Reports/IITR#declarerSignatureDate

Summary | JSON | XML | RDF | XBRL

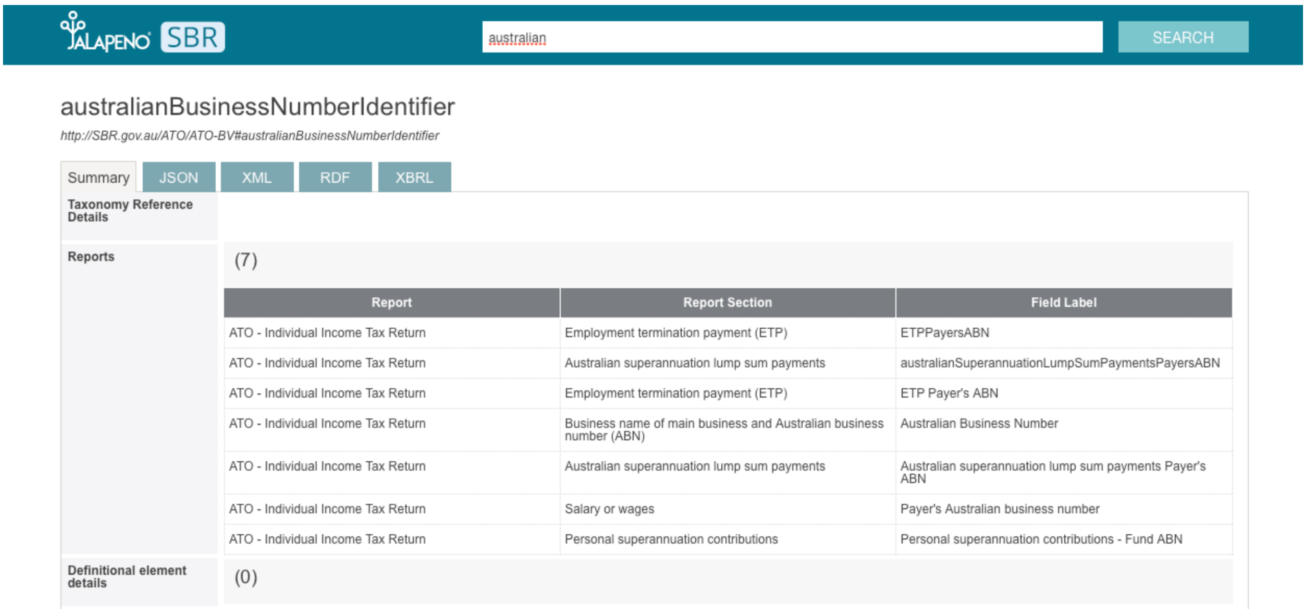
Reference Details	
Business Definition	Determines the date of the signature that ensures the acceptance of the terms of the Declaration
Taxonomy Element Name	Declaration.Signature.Date
Business Guidance	If no date is present, report 00000000 in this field. The date reported must be the actual date on which the declaration was signed by the payee, NOT the date on which the payer/payee relationship commenced or any 'default' or 'dummy' date.
Element Details	
Report Definition	
Report Guidance	
Report Element Name	Declarer Signature date
Datatype name	monetaryItemType
Report Element	(1)

Form Element ID	TREFID	Source Ref	min Occurs	max Occurs	namespace Prefix	Period
IITR.0001543	4		1	1	pyin.02.00	Duration

Fig. 3.10 Summary description for example report element

Definitional Element Summary

The summary of a definitional element includes details of the reports, which utilise that element (see Fig 3.11).



JALAPENO SBR | australian | SEARCH

australianBusinessNumberIdentifier

http://SBR.gov.au/ATO/ATO-BV#australianBusinessNumberIdentifier

Summary | JSON | XML | RDF | XBRL

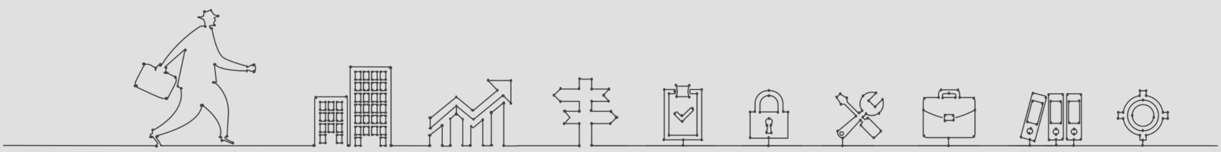
Taxonomy Reference Details

Reports (7)

Report	Report Section	Field Label
ATO - Individual Income Tax Return	Employment termination payment (ETP)	ETPPayersABN
ATO - Individual Income Tax Return	Australian superannuation lump sum payments	australianSuperannuationLumpSumPaymentsPayersABN
ATO - Individual Income Tax Return	Employment termination payment (ETP)	ETP Payer's ABN
ATO - Individual Income Tax Return	Business name of main business and Australian business number (ABN)	Australian Business Number
ATO - Individual Income Tax Return	Australian superannuation lump sum payments	Australian superannuation lump sum payments Payer's ABN
ATO - Individual Income Tax Return	Salary or wages	Payer's Australian business number
ATO - Individual Income Tax Return	Personal superannuation contributions	Personal superannuation contributions - Fund ABN

Definitional element details (0)

Fig. 3.11 Summary description for an example definitional element



Employer Reporting Domain

The scope of content selected to demonstrate the capabilities offered by a semantic technology platform is ‘Payroll’ within the domain of Employer Reporting.

Sample content was used as follows;

Individual Income Tax Return (IITR)	Provided by ATO (SBR Program). This is an example of a ‘report’, or message, generated in the fulfilment of a reporting obligation by a reporting business entity.
SBR Definitional Taxonomy	Maintained by ATO (SBR Program). (taxonomy-collaboration.sbr.gov.au) The definitional taxonomy is the ‘dictionary of terms’ managed within the SBR Architecture. Only the definitional elements used by the IITR report were included in the proof of concept.
SBR Payroll Reference Model	Provided by ATO (SBR Program). SBR Program document: *SBR_Information_Architecture_v04 – Martijn.docx* This is an example of a reference model describing a domain, to which reports in that domain should align.

In addition to the content above, further content was included to enrich the modelling.

APQC Process Classification Framework (PCF)	Source: apqc.org This is an open standard which defines a taxonomy of business processes. The processes themselves are not described. The PCF was originally intended to facilitate the benchmarking of processes between

	organisations.
HR-xml	Source: hropenstandards.org Reference model containing XML specifications enabling the standardisation of HR-related data exchange.

The proof of concept sought to demonstrate the potential benefits from the convergence of the above content.

- The modelling of report/message content and its relationship to definitional content
- A resource evolved by a business process task.
- Rendering of content as a conceptual model, as a relational model, and in serialised format
- Digestion of an industry reference model (presented as an ERD)

Resource Cell – Business View (Computational Independent Model)

Fig 3.12 below shows the concept 'Payslip' defined as part of the SBR Payroll Reference Model.

This view of a concept shows the following:

- A taxonomy (tree) of concepts with the selected concept highlighted.
- A graph showing the concept, its attributes and relations to other concepts.
- Tabular detail of the properties of the concept.

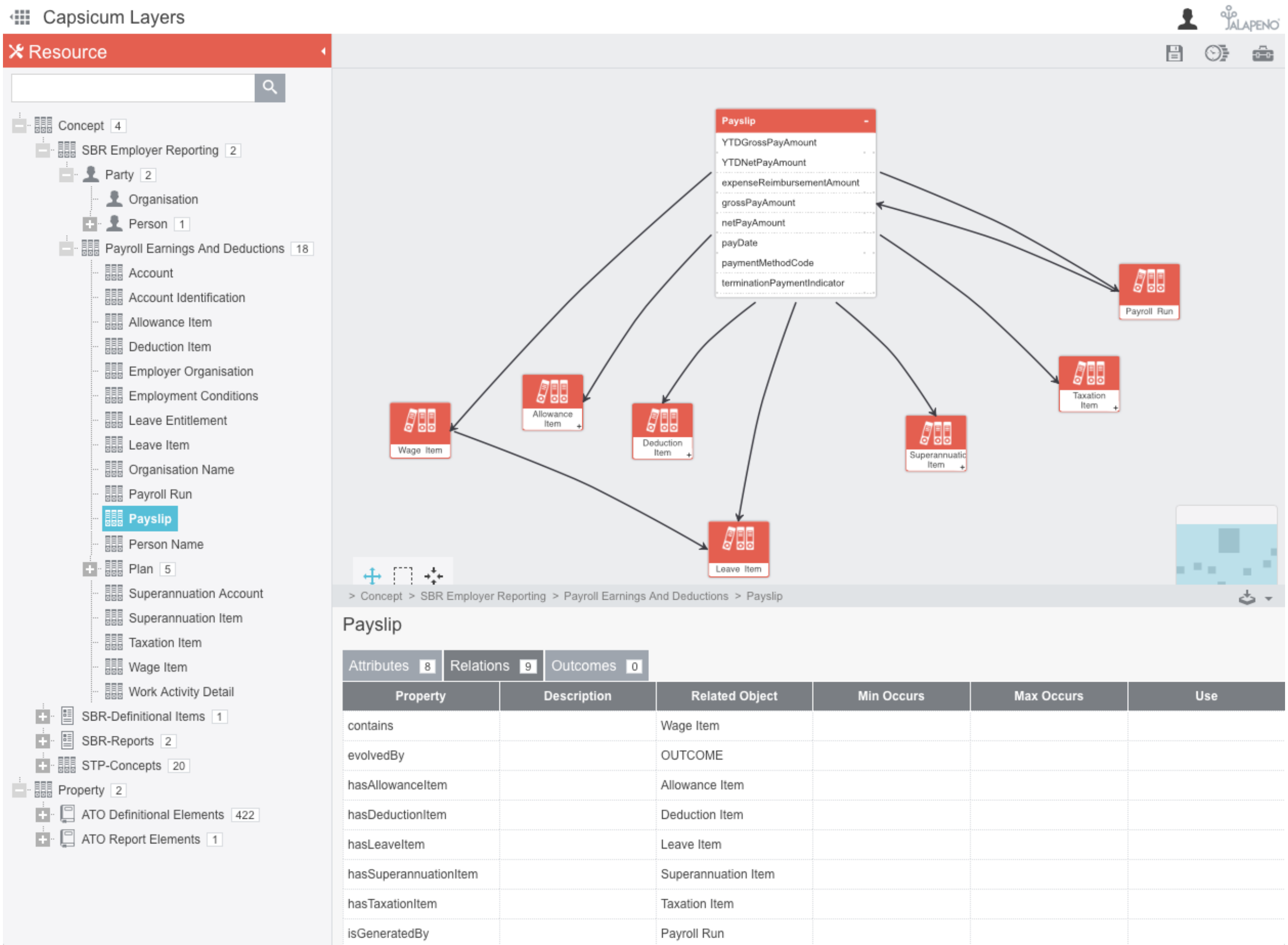


Fig. 3.12 Visualisation of concept 'Payslip'

Service Anatomy

The CAPSICUM Framework allows the granular modelling of business processes at task level as ‘Undertakings’. An Undertaking is;

- performed by a Role,
- evolves a Resource (concept),
- results in an Outcome (a change of state of a Resource)

In Fig. 3.13, the resource ‘Timesheet’ is evolved by the undertaking ‘SubmitTimesheet’ by the role ‘SubmitterTimesheet’ resulting in the outcome ‘SubmittedTimesheet’.

This structure, built around a verb (in this case ‘submit’), describing an undertaking is rendered as a ‘Service Anatomy’.

Fig. 3.13 shows the service anatomy for the undertaking ‘SubmitTimesheet’, modelled from the high-level Single Touch Payroll process definition.

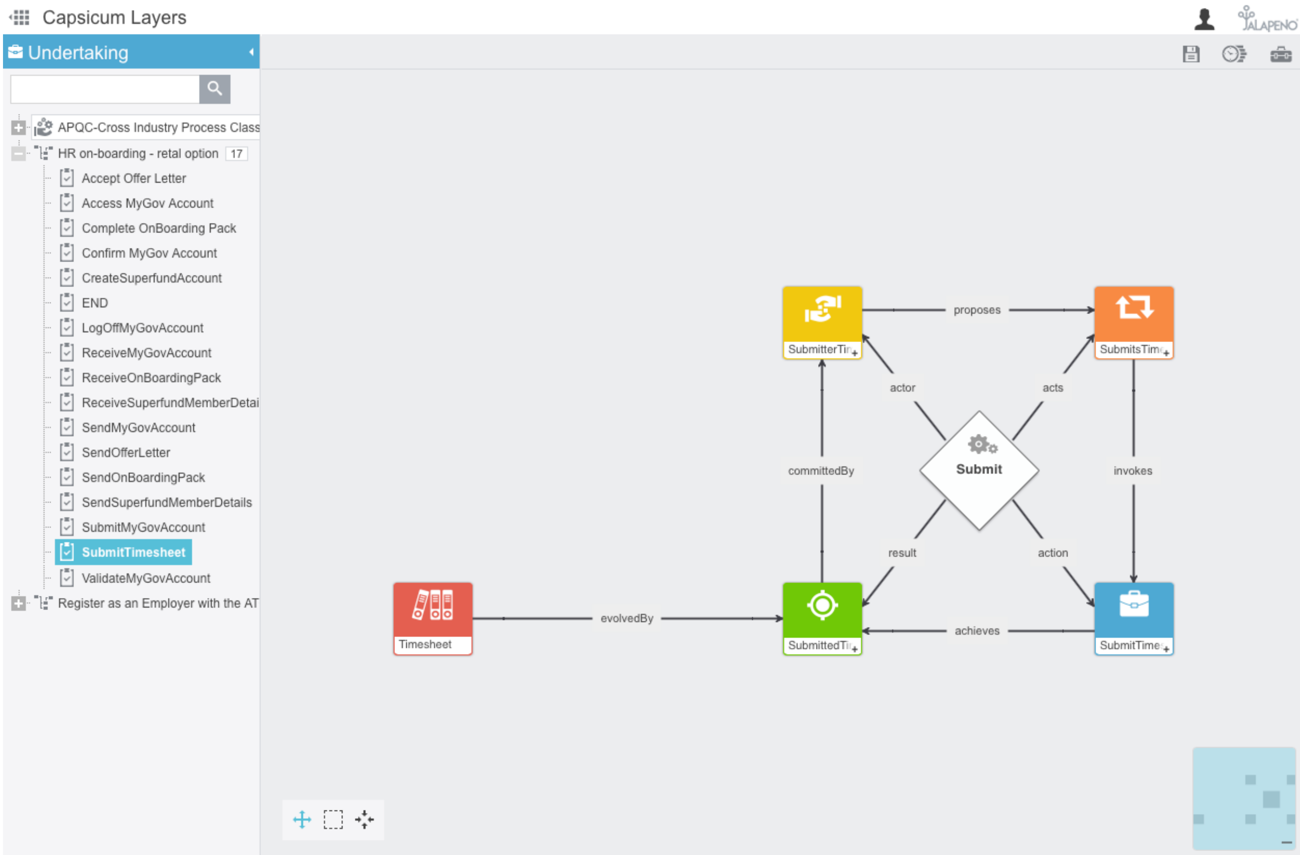


Fig. 3.13 Service anatomy for ‘SubmitTimesheet’

Technical View (Platform Independent Model) - DataModel

Concepts modelled in the Resource Cell in the Business View are rendered as a data model (or Entity Relationship Diagram) in the Technical View (DataModel Cell).

An example is shown in Fig. 3.14 below, which shows the concepts of the SBR Payroll Reference Model in ERD form.

The modelling tool allows the conversion of models between the Business View (conceptual models) and the Technical View (e.g. data models) in either direction.

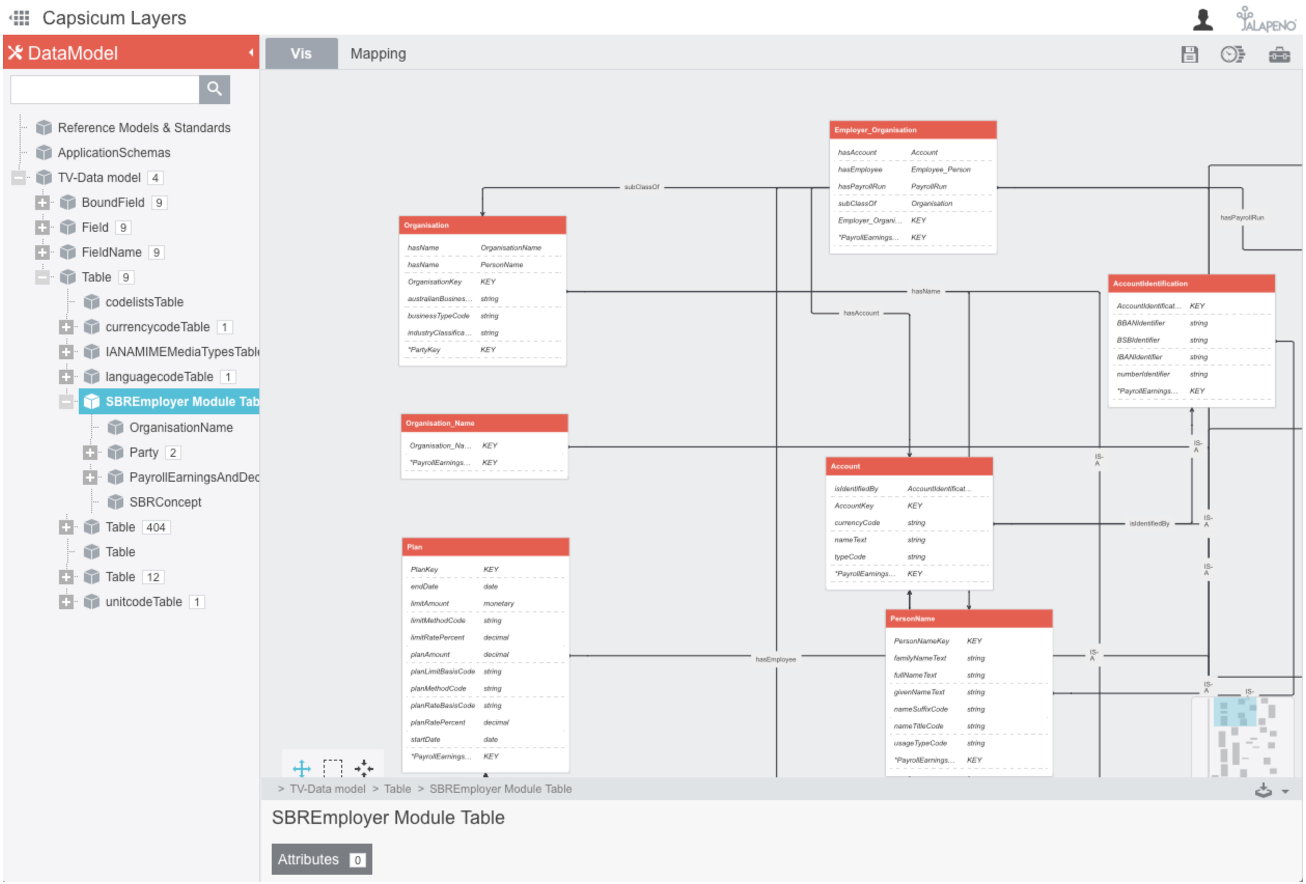
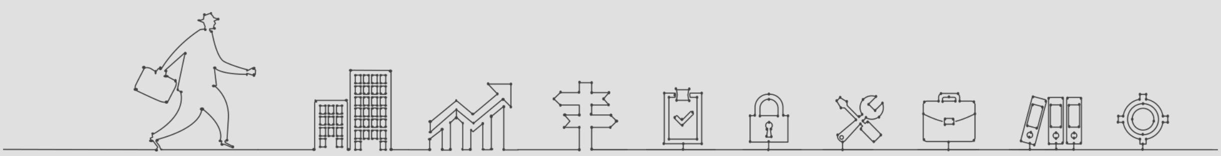


Fig. 3.14 Technical View, DataModel cell (Entity Relational Diagram) for SBR Employer Reference Model



Platform View (Platform Specific Model) – Schema

Fig. 3.15 below shows the concept ‘Payslip’ from the SBR Payroll Reference Model (shown in Fig. 3.12 in the Business View) rendered in a serialised format (in this case as an XSD schema).

Models can be converted between the Technical View (datamodels) and Platform View (schema) in either direction.

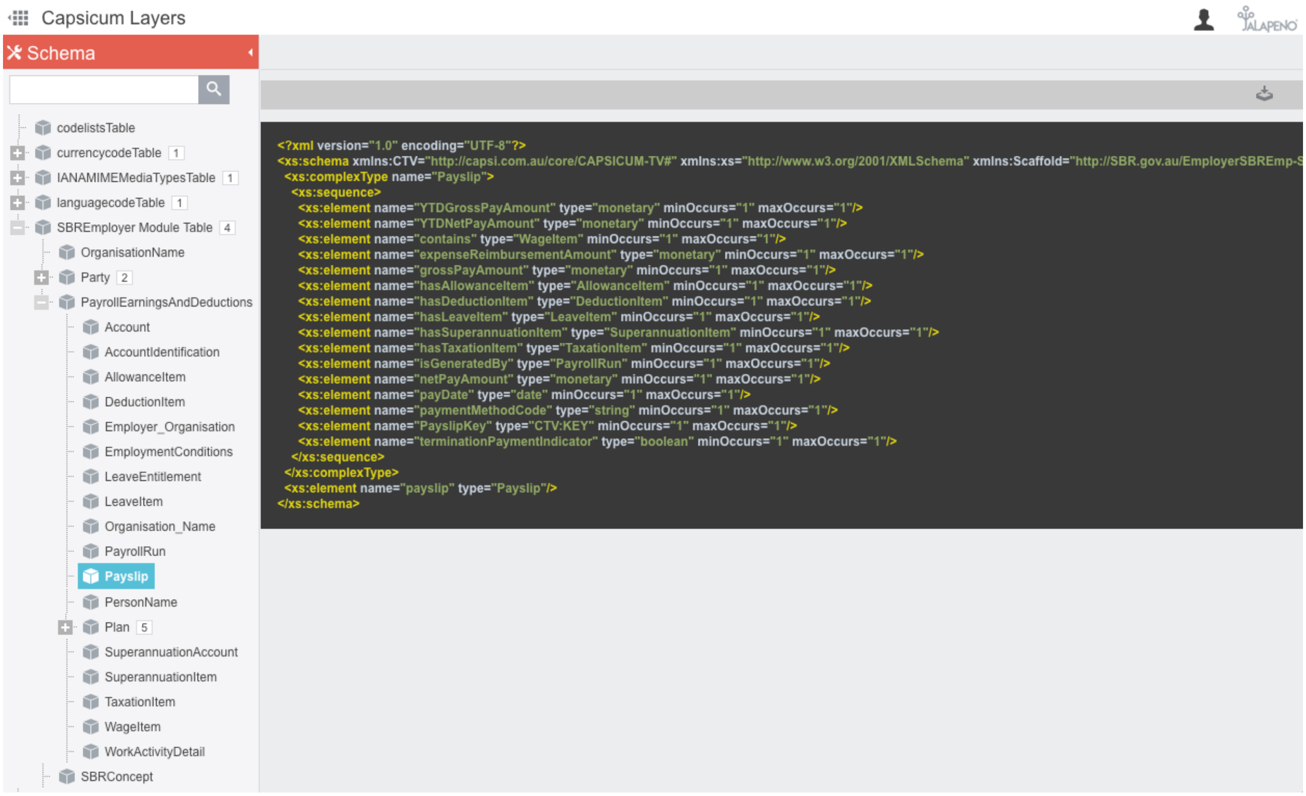


Fig. 3.15 ‘Payslip’ rendered as an XSD schema

PoC Findings

The two Proof-of-Concept exercises provided valuable opportunities to apply a semantic technology stack towards resolving some of the specific issues facing SBR, using real-world data and business problems.

Examples of the value derived and learning that was achieved from these exercises include:

- demonstration that XBRL based definitional elements can be readily converted to RDF and serialized from RDF into other common protocols (in this case XML and JSON);
- the effort involved in converting existing SBR assets to RDF is not a significant constraint and a one-off conversion of the existing assets from XBRL to RDF would be entirely reasonable and a very valuable exercise;
- publishing SBR content from RDF proved to be easy and cost-effective. The PoC RDF repository was easily incorporated into a model-driven prototype application of the Australian Reporting Dictionary. The prototype was of a high quality, was constructed in a matter of days, provided valuable additional content not available in the current ARD and could be readily exposed as a web-based, public-facing, query tool;
- graphic visualization of the SBR content facilitates greater understanding of the content and makes the assets more accessible to non-technical stakeholders;
- business processes can be semantically aligned with the supporting information models allowing for better understanding of the intersection of the information collected during a business process with contextual reporting obligations. There is great opportunity to reduce the regulatory reporting burden on employers by mapping their process information with the required information in the reporting standard;
- further opportunity exists to incorporate business rules into reporting standard providing further value to the reporting entities;
- RDF provides good opportunities for mapping the SBR models to other industry standards (in this case the mappings were to HR-XML, but this capability is similarly relevant to any other standard, for example the various health and financial industry standards that are already available in RDF);

Governance of a Semantic Model

An obvious question that arises from these recommendations is whether a semantic toolset provides any particular benefits or presents any particular challenges with regard to the ongoing governance of a semantic standard.

Some key areas worth discussing in this regard include:

Common Core Definitions Supporting Cross-Domain Requirements

An important characteristic of an RDF-based modeling approach is the ability to create a multi-layered, component-based model architecture. Effectively this allows you to factor your models in such a way that entire "graphs" can be imported into a model as an "overlay", extending the overall knowledge in the combined model whilst still preserving the integrity of each of the parts. The entire premise behind an Ontology is to develop a single, reusable conceptualization of a domain that can inform and be extended by any subsequent consumer of the Ontology.

In this way a reference model like SBR can be imported and aligned with an operational model of a business, rules repositories can be shared across business units, common party definitions can be re-used by different industry domains etc. It is quite typical for a domain model to comprise any number of distinct namespaces that reuse existing Ontologies that have been built as independent but extendable, modularized knowledge-bases.

SBR will find this a particularly useful feature for establishing whole-of-economy definitional elements that can be re-used across domains. Typical examples of this are Party models, where it is often desirable to define common standard conceptual representations of Contact Details and Addresses for Persons and Organizations which are applied consist provide greater consistency, it will ensure that knowledge about can be easily shared and new knowledge created through the combination of data. Common definitions supports "knowledge-integration" on the basis of:

> ***"if I know something about a thing and you know something about a thing, we will both know more if we can prove it is the same thing"***

Versioning and Release Strategies

One important challenge with maintaining an information model is the ongoing governance as the business vocabulary and the reporting environment inevitably change and the reporting landscape evolves.

There are clearly a variety of version control issues relating to the ongoing management of an information standard. It should be expected that the following minimum requirements would need to be addressed:

- The ability to extract a schema version from a modelling repository at any point in time via a manual schema "dump" and/or via a parametric web service interface;
- The ability to maintain immutable baseline version snapshots which are dynamically accessible by an external system for the purposes of mapping incoming messages;
- A consistent method for uniquely identifying schema versions through a semantic version numbering approach;
- The ability to maintain a set of mapping values from the relevant source and target messages and vice versa;
- The ability to maintain an audit trail, log changes, manage clashes and roll-back changes (achieved through a 3rd party tool - GIT);
- A mechanism for exposing version details to any systems that need to know, minimising the manual steps involved in publishing and consuming schema updates to reduce the risk of errors and inconsistencies.

Since it will clearly be required that multiple authors may need to edit the standard at the same time, additional version control features for supporting collaborative updating of a model includes:

- The capability for collaborative, multi-user updates to the schema models between published versions;
- The ability to timestamp each update to a model element and to roll back to a version of the models at any discrete point in time;
- The capability for modellers to create flexible workflow processes to manage the review and approval of model updates and to track the status of model updates through a user-defined set of governance statuses (available through 3rd party integration).
- The ability to maintain an audit trail of each model update by the timestamp, user ID and the specific change that was made to the model;

Semantic tooling implemented on bi-temporal datastores allow time-based information to be continuously associated with entries to the database which effectively records a complete history of all user actions. When a new model is created for the first time a "Day-Zero" timestamp is associated with the initial model load. From that time on, as a user updates the model, each write to the triple-store is automatically time-stamped.

Existing models can be loaded from the server file system without re-doing a day zero load. Once loaded, the server can be reset to any point in its history since day zero. A temporal management page shows all relevant temporal boundaries since day zero, and offers the user the opportunity to select a model version from any of these contexts.

It is evident that the SBR models will be in continuous evolution and that modellers will need to make updates and changes to the standards. Once an initial baseline version of a standard has been established, it will become extremely important to implement model governance controls to monitor and manage any changes that are proposed. Procedures will need to be established for tracking, reviewing and approving changes to the model and defining how and when to publish a new version.