The Case for the Standardised Semantic Information Model

Version A 13 December 2018

Contents

Preamble

This document sets out why the Standardised Semantic Information Model (SSIM) is needed, the advantages it offers versus alternatives, how it will be used, and how it will gain adoption.

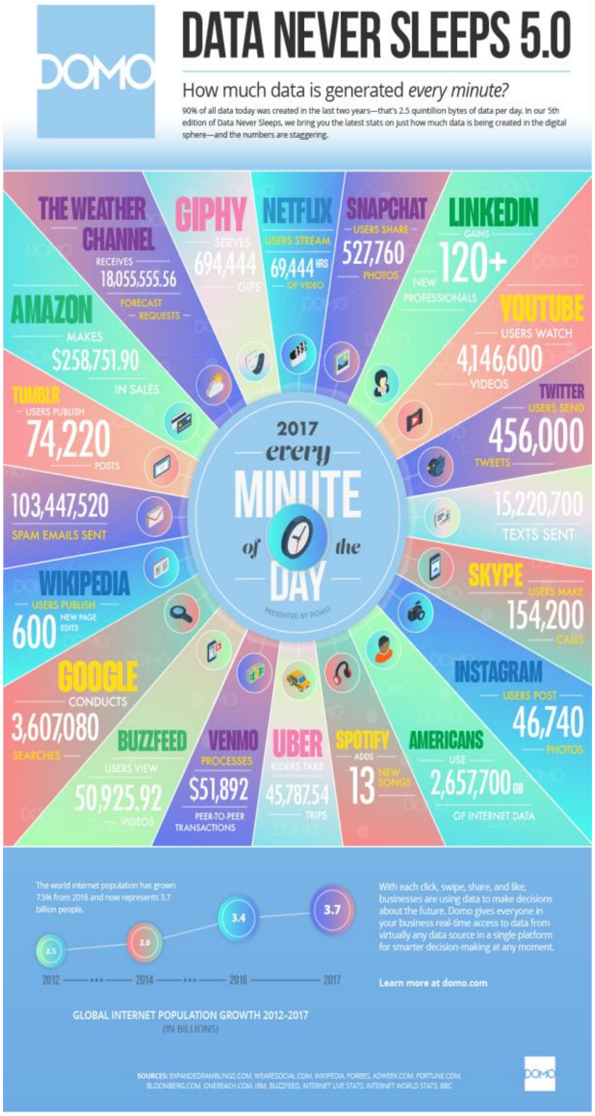
Because SSIM is currently a work in progress, everything in this document is subject to change.

Greater technical details are provided in the separate document "Standardised Semantic Information Model (SSIM)" in SSIM.docx".

1 The Standardised Semantic Data Need

The 21st century is the digital century, the age of the internet, artificial intelligence, software, blockchains and cryptocurrencies all driven by data affecting virtually every aspect of life. The flood of data is "truly mind-boggling. There are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT). Over the last two years alone 90 percent of the data in the world was generated."[1]

Data never sleeps tries to present an idea of the flood of data as follows:



But that diagram is far from complete. It leaves out business data, business transactions, research data, government data, charity data, IoT data, and personal plus business communication.

Further, all that data is mostly in incompatible formats, stored in data silos that aren't easily accessible, unless to hackers! Yes, data as it is stored and managed poses great security risks, as shown by the Marriott hack revealed in Nov 2018 which affected 500 million people back to 2014.

Information overload predicted by the 1970 book Future Shock[2] is upon us.

---

[1] Forbes: How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read

[2] Future Shock is a 1970 book by the futurists Alvin and Heidi Toffler, in which the authors define the term "future shock" as a certain psychological state of individuals and entire societies. Their shortest definition for the term is a personal perception of "too much change in too short a period of time"

Yet in some ways the digital transformation is only getting started. IDC[3] predicts that "by 2022, over 60% of global GDP will be digitized with growth in every industry driven by digitally-enhanced offerings, operations, and relationships."

That IDC report also predicted that "From 2018 to 2023, with new tools/platforms, more developers, agile methods, and lots of code reuse, 500 million new logical apps will be created, equal to the number built over the past 40 years." The term "app" is meant broadly here to cover software applications of all kinds, including mobile apps, blockchain dapps (distributed apps), web apps, and traditional business/computer applications.

Those 500 million new apps will all generate or use data in some form, and many of them will involve the "relationships" mentioned in the first prediction.

How can a person, organisation, or app cope?

The only solution other than returning to a pre-digital era which is inconceivable unless from total world collapse, is to improve the technology to manage the flood.

That requires four things:

- Networks and storage systems which can handle the volumes
- Improved security so that what is private stays private with what is public or public by need can be confidently controlled by the owner/creator of the data
- Machine (computer) understandable data so that programs or apps can know what the data is about, optionally where relevant at a specific date/time, and automate searching it, reporting on it, combining it, processing it, whatever, without a human needing to wade through the flood. The short way to say that is that the data needs to be "machine readable semantic" where data can be "understood" by software.
- Standardisation so that machine understanding is consistent within context i.e. so that "apple" is a fruit in one context and the name of a company in another, or that rent is a rent expense to a tenant but rental income to the landlord, even if both are referring to the same number e.g. €1,000 in an invoice or receipt.

The benefits which will flow from success are huge, with new opportunities, and the survival of businesses which would otherwise fail. Multiple trillions per annum are involved. Success will even be a part of the survival of humanity on earth as data is a key part of understanding what is happening with climate change and managing it. What is the price to put on that?

Whereas the cost to the world of failure is also huge ranging from trillions of dollars per annum in quantifiable costs to playing a part in the extinction of humanity. Examples of the quantifiable costs are world company fraud (misuse/manipulation of data) estimated at $4 trillion[4] per annum, and reconciliation of incompatible financial records adding another $780 billion[5].

The first two bullet points (capacity and security) can be handled thanks to improving hardware and system software, including blockchains, with Pacio's platform playing a role.

Many have tried to achieve the next two bullet points (semantic standardised data), with a lot of effort expended, and some success, as described in the next section, but full victory remains elusive.

This document presents Pacio's proposal for achieving standardised semantic data success.

---

[3] IDC Report 10 predictions for how the tech industry will change in 2019 and beyond

[4] 2018 REPORT TO THE NATIONS | 2018 GLOBAL STUDY ON OCCUPATIONAL FRAUD AND ABUSE
Page 8: To place their estimate in context, if the 5% loss estimate were applied to the 2017 estimated Gross World Product of USD 79.6 trillion, it would result in a projected total global fraud loss of nearly USD 4 trillion

[5] XBRL International Newsletter 13 April 2018 IFAC: International regulatory divergence costs $780B

2 Semantic Data Initiatives

2.1 The Semantic Web

One hope for achieving semantic data was the so called Semantic Web or Web 3.0 where information on the web was to be structured or classified "semantically" so that it could be accessed and used according to content, so, ideally, allowing people to interact in mutually beneficial ways without needing to know or trust each other. "The great promise of the Semantic Web was that it would be readable not just by humans but also by machines. Pages on the web would be meaningful to software programs — they would have semantics — allowing programs to interact with the web the same way that people do. Programs could exchange data across the Semantic Web without having to be explicitly engineered to talk to each other."[6]

The World Wide Web Consortium (W3C) headed by Sir Tim Berners Lee, the inventor of the world wide web, put in a massive effort generating specifications and standards to enable development of the Semantic Web. This includes the Web Ontology Language (OWL) as a family of knowledge representation languages for authoring ontologies, where ontologies are a formal way to describe the structure of knowledge for various domains.

Numerous other organisations have contributed also, including to development of the so called web 3 stack, yet, after almost 20 years, the full semantic web vision has not eventuated, and instead we have centralised control by a few large corporations.

Sir Tim himself says "I Was Devastated"by what the web has become – dominated by centralised giants, the FANGs (Facebook, Apple, Amazon, Netflix and Google) rather than the decentralised equalising semantic web for all that he had envisaged. He has taken leave of absence from W3C to head a project named Solid which "aims to radically change the way Web applications work today, resulting in true data ownership as well as improved privacy."

Whatever Happened to the Semantic Web? lists a number of hypotheses as to why the web or internet has evolved to what we know today rather than the semantic web, and describes the four phases of the long but ultimately unsuccessful effort. It is a good read for anyone interested in the subject. It ends with:

> So the problems that confronted the Semantic Web were more numerous and profound than just "XML sucks." All the same, it's hard to believe that the Semantic Web is truly dead and gone. Some of the particular technologies that the W3C dreamed up in the early 2000s may not have a future, but the decentralized vision of the web that Tim Berners-Lee and his follow researchers described in Scientific American is too compelling to simply disappear. Imagine a web where, rather than filling out the same tedious form every time you register for a service, you were somehow able to authorize services to get that information from your own website. Imagine a Facebook that keeps your list of friends, hosted on your own website, up-to-date, rather than vice-versa. Basically, the Semantic Web was going to be a web where everyone gets to have their own personal REST API, whether they know the first thing about computers or not. Conceived of that way, it's easy to see why the Semantic Web hasn't yet been realized. There are so many engineering and security issues to sort out between here and there. But it's also easy to see why the dream of the Semantic Web seduced so many people.

That conclusion contains a clue as to how the ideal might be achieved since one of the key objectives of blockchain technology based systems is to return control of data to its owners/creators.

---

[6] From Whatever Happened to the Semantic Web?

Perhaps in hindsight this article from 2001 Metacrap: Putting the torch to seven straw-men of the meta-utopia, mentioned in the "Whatever Happened to the Semantic Web?" article was prescient when it pointed that creating metadata (data to describe data) is hard because of self interest and conflicts of interest. SEO (Search Engine Optimisation) tricks to game the system, fake news, and the interminable arguments about some Wikipedia pages show just how hard it is, ultimately rendering the whole effort futile.

Though some are still trying to reach the semantic web utopia, for example Schema.org, and Sir Tim's new effort, it is clear that the semantic web effort has lost momentum, and even if not actually buried yet, does not look likely to ever provide the full answer to the need for semantic data at scale.

2.2 Knowledge Graphs and Public Datasets

One form of a collection of knowledge resulting from or inspired by the semantic web hope is called a knowledge graph. Google use knowledge graphs as does Facebook in social graphs and its Open Graph.

Freebase which launched in 2007 was a pioneering example: "Freebase was a large collaborative knowledge base consisting of data composed mainly by its community members." As of January 2014, Freebase had approximately 44 million topics and 2.4 billion facts.

Although the Freebase data is still available, Freebase is no longer active as it was acquired by Google in 2010 and shut down in 2014.

The Freebase data was used as part of Google's own knowledge graph, and also in the dataCommons project started by Google. The dataCommons page describes the problems well:

> Publicly available data from open sources (i.e. census.gov, NOAA, data.gov etc) are a vital resource for students and researchers in a variety of disciplines. Unfortunately, processing these datasets is often tedious and cumbersome. Organizations follow distinctive practices for codifying datasets. Combining data from different sources requires mapping common entities (city, county, etc) and resolving different types of keys/identifiers. This process is time consuming and can increase the likelihood for methodological errors.
>
> dataCommons attempts to synthesize a single Knowledge Graph from these different data sources. It links references to the same entities (such as cities, counties, organizations, etc.) across different datasets to nodes on the graph, so that users can access data about a particular entity aggregated from different sources. Like the Web, the dataCommons graph is open - any user can contribute datasets or build applications powered by the graph. In the long term, we hope the data contained within the dataCommons graph will be useful to students and researchers across different disciplines. Though we've already "jump-started" the graph with data from publicly available sources (Wikipedia, US Census, FBI, State election boards, etc), we encourage you to join and contribute.

Then there is DBpedia which "is a crowd-sourced community effort to extract structured content from the information created in various Wikimedia projects. This structured information resembles an open knowledge graph (OKG) which is available for everyone on the Web. A knowledge graph is a special kind of database which stores knowledge in a machine-readable form and provides a means for information to be collected, organised, shared, searched and utilised."

The English version of the DBpedia knowledge base describes 4.58 million things, out of which 4.22 million are classified in a consistent ontology, including 1,445,000 persons, 735,000 places (including 478,000 populated places), 411,000 creative works (including 123,000 music albums, 87,000 films and 19,000 video games), 241,000 organizations (including 58,000 companies and 49,000 educational institutions), 251,000 species and 6,000 diseases. In addition, DBpedia provides localised versions of DBpedia in 125 languages.

DBpedia can be queried online and has been used this way by organisations such as the BBC, NY Times, and IBM, which employed DBpedia as a knowledge source for IBM Watson, the Jeopardy-winning artificial intelligence system.

Other knowledge graph or knowledge graph like sources are Wikidata (and at Wikipedia) plus the AWS Public Dataset Program.

Clearly knowledge graphs and public datasets have a role to play and will continue to be developed and used.

But are they widely enough used, standardised, date and time specific, and reliable enough for use by the 500 million apps coming through in the next five years to depend on them?

No in many ways. Standardisation is often doubtful or different across graphs. Mostly data which changes with time e.g. exchange rates is not handled. Nor even are less frequent time changes covered in any standard way which would of use to an app e.g. Britain in 2018 is an EU member, but not after March 2019 if Brexit proceeds. In short, these public knowledge bases have not been designed to service the needs of apps or blockchain oracles (knowledge serving apps), for standardised, time specific, high reliability semantic data at scale. (When writing this on 2018.12.08 all DBpedia sample views were broken.)

Available knowledge graph and public dataset repositories of knowledge can provide inputs to a standardised reliable system of semantic knowledge for use by apps at scale, but not be that system.

2.3 Other Ontology Efforts

There has been much academic and business work to produce ontologies, usually specific to certain knowledge domains.

An old (2006) Boeing paper Ontologies Ontologies Everywhere – but Who Knows What to Think? Is still relevant as there remains no clear option for wide scale use by apps.

Corporate ontologies are used by larger entities as described in the book “Semantic Applications | Methodology, Technology, Corporate Use”[7].

Most such ontologies result in the data described by them remaining in corporate silos, but one which is at least one industry wide is Financial Industry Business Ontology (FIBO™) developed by the EDM Council as “the open semantic standard for the financial industry - a business conceptual model developed by our members of how all financial instruments, business entities and processes work in the financial industry.”

Such efforts are how some entities and industries are coping. They do not, however, provide a general solution, nor are they available or suitable for the bulk of apps – the 500 million apps of the next 5 years predicted by IDC.

As for knowledge graphs, “other ontology efforts” can provide inputs to a standardised reliable system of semantic knowledge for use by apps at scale, but not be that system.

2.4 Data Description Languages and Protocols

Numerous data description languages and protocols have been developed, mainly for IT (Information Technology) use generally, but also in part spurred by the pursuit of semantic data. These include:

- XML (eXtensible Markup Language) which is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. W3C specifications—all of them free open standards—define it.

---

[7] Springer 2018 “Semantic Applications | Methodology, Technology, Corporate Use” Editors: Hoppe, Thomas, Humm, Bernhard, Reibold, Anatol (Eds.)

Although the design of XML focuses on documents, the language is widely used for the representation of arbitrary data structures such as those used in web services.

- JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write, and also easy for machines to parse and generate. JSON has become the most popular data-interchange format – passing XML in use – largely because of its simplicity, with the complete language being defined on one page rather than in a lengthy set of specifications. Originally it did not have a scheme option available, unlike XML, but in recent years that deficiency has been corrected.

  JSON-LD with the "LD" standing for Linked Data is an extension of JSON to provide semantic information similarly to RDF below. There is also a binary form of JSON used for more efficient data storage/transfer. (Normal JSON is text based and verbose, though not as verbose as XML.)

  The Whatever Happened to the Semantic Web? article mentioned earlier in *The Standardised Semantic Data Need* section mentions the rise of JSON and JSON-LD vs XML. One of its postulates is that the Semantic Web might have achieved greater success if JSON had arrived earlier and been used by the W3C instead of XML.

- **UBL** or International Standard ISO/IEC 19845 "specifies the OASIS Universal Business Language (UBL), which defines a generic XML interchange format for business documents that can be restricted or extended to meet the requirements of particular industries.

- **Open EDI** or International Standard ISO/IEC 14662:2010 "describes, through two perspectives of business transactions, significant aspects relevant to the interoperability of information technology systems used by Open EDI Parties engaging in Open-edi.

- **RDF** or Resource Description Framework is the W3C standard model for data interchange on the Web as part of OWL – the Web Ontology Language. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

  Though RDF has achieved some acceptance it is complicated and verbose, and has failed to take over the world, as discussed in *The Semantic Web* section above.

- **Ocean** is A Decentralized Data Exchange Protocol to Unlock Data for AI. From the site: Ocean Protocol is supported by a Singapore based non-profit foundation (Ocean Protocol Foundation Ltd), whose mandate is to ensure open access to the protocol and platform, provide data governance, encourage the network ecosystem growth and take measures to ensure that the platform becomes ever more decentralized with time.

- CBOR (RFC 7049 Concise Binary Object Representation) and other protocols or proposals including ION (there are two IONs, one by Amazon and one preceding the Amazon proposal), Protobuf, and MessagePack.

These protocols can form parts of possible solutions or systems, but none of them provides a complete solution to the need for standardised semantic data at scale.

2.5 XBRL (eXtensible Business Reporting Language)

XBRL (eXtensible Business Reporting Language) is a standard language for reporting on or comparing business information built upon XML. XBRL has been adopted by most developed countries for the benefit of shareholders and government to standardise business and financial reporting to increase the transparency and accessibility of business information by using a uniform format. It has become the effective world standard for business reporting.

XBRL defines context and classifies the data in a financial report so it is semantic. It goes further to define the structure of reports in definition, presentation, and calculation forms via Taxonomies. Hundreds of taxonomies now exist for the accounting standards and jurisdictions of the world. XBRL reporting for businesses is mandatory in some countries including the UK for all companies, and in the US for public entities.

XBRL may be considered to be the most successful of all semantic data initiatives - it has momentum and widespread use.

Given the success of XBRL, why is there a need to consider anything else?

The Pacio view is that XBRL is not a complete solution for the standardised semantic data at scale need. Specifically:

XBRL is, as its name says, a reporting language, and thus not ideal for data storage definition and transactional data interchange purposes. Its creator, Charles Hoffman, argues that as the existing (financial reporting) standard it could be put to use at all levels, including for an XBRL equivalent of the web 3 stack, but the Pacio view is that for the revolution in progress with the change to the blockchain era, it is undesirable or inappropriate to coerce a complicated verbose text based protocol into doing something for which it was not designed, when the changes in progress provide the opportunity to do so much better. This is especially the case for blockchain use where efficiency matters.

XBRL is specific to the financial reporting domain – it was not intended to be general.

Despite covering only financial reporting, XBRL taxonomies have become large and complicated. The full 2018 US GAAP taxonomy used for company reporting in the USA defines 19,494 elements (reporting concepts) and the relationships between them via 385,595 XML nodes in 547 files. 574 elements were deprecated in the 2018 taxonomy. Deprecated elements remain in the taxonomy for two annual updates. Accordingly, 441 such elements from the 2015 taxonomy were deleted from the 2018 one. Deprecated elements greatly complicate comparisons over time.

The IFRS (International Financial Reporting Standards®) taxonomy used for financial reporting in much of the world other than the USA, is smaller, consisting of “only” 88,688 XML nodes in 2018, yet that remains a lot.

One of the benefits of text based protocols such as XBRL is that data is human as well as machine readable. However, some XBRL taxonomy names or tags have become so long as to render that of little value. The IFRS taxonomy in particular uses many long names. The longest in the 2018 taxonomy is 234 characters long: “IncreaseDecreaseThroughReclassificationAdjustmentsInPeriodReserveOfGainsAndLossesOnFinancialAssetsMeasuredAtFairValueThroughOtherComprehensiveIncomeRelatedToInsuranceContractsToWhichParagraphsC18bC19bC24bAndC24cOfIFRS17HaveBeenApplied” which is not exactly human friendly!

Additional financial reporting requirements such as ESG (Environmental Social Governance) which are expected to become mandatory in some jurisdictions over the next few years, will further greatly complicate the XBRL taxonomies.

These XBRL taxonomies have a compliance (regulatory) reporting focus, and are not ideal for management reporting, yet it is for business improvement in real time where semantic data can be especially useful.

As Dr Keith Cleland, one of the Pacio founders, points out a 'missing link' in management financial reporting is productivity identified as units of output v units available. Whilst management accounts have this information expressed as variances of labour and or material from target or budget, there is no cross pollination of what amounts to vital management information with typical financial reports, be they monthly, quarterly or annually. In any event, such management accounting data is non-existent in current accounting software such as SAP, Intuit, Sage, Xero, MYOB, used by SMEs whick make up more than 90% of reporting businesses.

The work of Keith and his partner Trevor Watters, another Pacio c-founder, has proved that businesses applying the TARI®[8] concepts built into Pacio achieve a minimum of 10% increase in productivity with resulting bottom-line improvement. That is huge for any economy, and for the world. XBRL taxonomies as they exist today do not address this need. Catering for it would further complicate them and be unlikely to result in actual use by applications actually used in business.

Some, such as Charles Hoffman, argue that taxonomy complexity doesn't really matter given appropriate software to hide the details, and assist in ensuring consistency. Charles is doing good work to make this possible as described in Intelligent XBRL-based Digital Financial Reporting and Accounting Process Automation XBRL Application Profile, an efficient, effective, safe, and reliable approach to implementing accounting process automation leveraging the XBRL global standard technical syntax. (The second document acknowledges the SSIM proposal with the words "proprietary approaches to representing semantics such as SSIM", though Pacio is making SSIM open, not proprietary.)

While admiring the effort, commitment, and energy displayed by the XBRL developers, and those building and using XBRL taxonomies, the Pacio view given all of the above is that XBRL is not the best system to use for building the future of standardised semantic data at scale.

Rather, Pacio proposes that it is time to consider an alternative to XBRL (and the other initiatives) that is better suited to the task, though any such system will need to be able to interact with XBRL and other systems for compatibility and continuity purposes.

3 The Standardised Semantic Information Model (SSIM) Proposal

Pacio proposes the Standardised Semantic Information Model (SSIM) as a method of storing, categorising, and reporting on data in a standardised, semantic way which is powerful, general, open ended, easy to understand and use at a user level, while being efficient for scaling to large scale use. SSIM is proposed as being better suited than are the current alternatives discussed in the previous section to the task ahead of managing the flood of data in the world to allow apps to reap the benefits of that data and to pass them through to users.

SSIM is being developed by Pacio and will be used by Pacio but is intended to be an open source standard which can be used independently of Pacio.

SSIM applies to all data, but the initial development focus is on business data.

SSIM takes a bottom up approach to providing standardisation so that raw data categorised by SSIM can be recast or aggregated to suit any desired reporting requirement. In this way SSIM applies from a single raw transaction such as the sale of one can of beans all the way through to financial statements or <IR> integrated reports and ESG (Environmental Social Governance) reporting according to the desired accounting standard target, or targets e.g. US GAAP and IFRS, with the whole process fully automated from one end to the other.

SSIM provides semantic information by a flexible data description system based on many directories of facts, knowledge graphs, and public datasets e.g. countries, currencies, languages, roles etc, with support for multiple human languages and jurisdictions, plus reporting objects structured according to ontologies.

Any single piece of data can be fully described by a single 64 bit (8 byte) number called a SSIM Id or SID. SIDs are totally flexible yet efficient for blockchain/database use with their fixed and small size. No long or variable length tags are involved at the transaction and data storage levels.

---

[8] TARI® or Target Average Rate Index is a productivity measurement and business improvement methodology built into Pacio, and described in the Pacio White Paper.

SSIM covers all storage, classification, and reporting needs, but will also interact with or interface with other systems or data description languages or protocols such as XBRL, ODI, UBL, Open EDI, RDF, Ocean etc as required.

Users of apps making use of SSIM will not need to know anything about the underlying details. In most cases apps will be able to classify data themselves. In the few cases where a manual selection might be required, an app will be able to present a selection list for a user to choose from.

The facts directories, which are a major part of SSIM, are intuitive and simple to use.

Ontologies and other parts of SSIM are more complex, but modelling a complex world inevitably requires some degree of complexity. However, the complexity has been made easier to understand and work with than in other systems, by avoiding arcane mathematics and jargon. Spreadsheets are used as the visualisation and development tool – a tool familiar to most people who are likely to use SSIM. (The spreadsheets do not present an integrity danger – their inputs are checked for validity and integrity. They are just a familiar tool, and never used directly.)

All SSIM components will be open source and public, developed and maintained through a Pacio moderated folksonomy[9] process. They will provide the equivalent of many currently scattered non-integrated directories, knowledge graphs, and ontologies/taxonomies.

The following sections describe the details.

[9] Folksonomy is a user driven system of classifying and organizing online content. It was used by Freebase, a large (1.9 Billion triple) public knowledge base prior to its acquisition by Google.

4 SSIM Basics

SSIM provides a precise and concise way to describe any item of data by means of a single number called a SSIM Id or SID. Even though the description of an item in words might be lengthy e.g. "International Business Corporation incorporated in Saint Lucia, which is a member of the Organisation of Eastern Caribbean States, and a member of the Eastern Caribbean Currency Union", the SID would be just 8 bytes long, the same length as a SID for "rent".

The avoidance of lengthy and variable length tags helps with blockchain and database efficiency.

The SID can be used in searches, and by apps to identify data in a standard way, semantically, and accurately.

The precise definition of the descriptions making up a SID improve reporting and search accuracy.

To achieve this, SSIM at its starting or basic level uses four components:

- Digital Ids or DigIds to identify the person, entity, app/dapp involved
- A Data Type to define the structure or nature of the data being described
- Directories of facts to be used to semantically describe a data item, with as many facts being used as is needed to fully describe it, in a process similar to tagging, in most cases done automatically by the app involved
- The resulting SSIM Id or SID which combines all the above references into a single number

For business/financial applications SSIM goes further, but just this basic start is powerful in its own right. SSIM at this level can be used independently of Pacio.

Descriptions of the four components follow.

4.1 Digital Id or DigId

People, entities, and apps using SSIM need to be identified uniquely. SSIM and Pacio are not about anonymous data, but rather ensuring that people or entities have ownership and control over their data, and to profit from it, not others such as the infamous FANGs (Facebook, Apple, Amazon, Netflix and Google). This does require that the creators of data be identifiable. It does NOT mean that the data has to be public. Whether the data is public or private is a choice of the creator.

For business applications especially a data system like SSIM needs to be able to uniquely identify each participant. This is needed to correctly handle inter entity transactions. Example: ACME ltd sends an invoice to Widget Inc. Only when both entities are uniquely and irrevocably identified in the system can fraudulent invoicing be avoided. Company names can change or conflict. An Entity Id is intended to be unique and permanent – essential for efficient and safe business.

Pacio takes the process a step further than others by also identifying the app or dapp which was involved in creating a piece of data. That improves security and auditability.

Pacio will provide a digital id or digid service to identify people, entities, apps, and dapps which works in conjunction with other services:

- Open source technology and standards as listed in Decentralized Digital Identities and Blockchain:
    - Decentralized Identity Foundation (DIF)
    - Decentralized Identifiers (DIDs) – a W3C spec that defines a common document format for describing the state of a Decentralized Identifier
    - Identity Hubs – an encrypted identity datastore that features message/intent relay, attestation handling, and identity-specific compute endpoints.

  - Universal DID Resolver – a server that resolves DIDs across blockchains
  - Verifiable Credentials – a W3C spec that defines a document format for encoding DID-based attestations.

- Existing digital identity participants:
  Civic, DID (Decentralized ID), Essentia, Estonia, Legal Entity Identifier (LEI), OpenID, Persona, uPort, the United Nations ID2020 programme, VeriMe
- Other initiatives that gain traction

As with all aspects of Pacio, the Digital Identity service will be open, and designed to readily accommodate new standards or services as they become available.

DigIds will be used by Pacio but could also be used by non-Pacio systems as DigIds are not specific to Pacio.

4.2 Data Types Directory

In a standardised semantic environment, every item of data needs to be classified, whether that be a knowledge base entry about the universe, or a business record (invoice, purchase order, journal etc). Classification has two main attributes, what the item is in a generic sense (number, word, file etc), and a description of what the content is about. Example: the item could be "currency number with 2 places of decimals" for which the description is "USD, rent". The Data Types Directory defines the possible types of data items to be classified by SSIM.

Pacio will create and maintain a public directory of data types for data to be described by SSIM. The data types directory will be extended as needed in a Pacio moderated folksonomy[9] like process.

This directory will help with standardisation by defining data types in one place for use by all SSIM components, including the more complex ones described in following sections, without those components needing to repeat the definitions for the data types they use.

A data type can be:

- Null or nothing in the special case where the fact is the data e.g. a country, with the formatting for that fact being the country name in the target language, or the 3 letter code or the 2 letter code
- a single item such as a number, a currency/money item, a datetime, or some text etc
- larger items also such as document, image, video, file … any digital thing.
- a number of items e.g. a money amount, a quantity number, an activity units number, a barcode (inventory) reference etc as needed.
- fields repeated n times for periodic data as for financial data in monthly or other periods, where the periods and dates of the periods are an entity property
- optional restrictions e.g. number must be positive, or must have 6 digits etc according to rules or patterns, but done without using complicated regular expressions that most people wouldn't understand
- formatting information where relevant, potentially varied according to country or jurisdiction and/or human language.

A data type is referenced by an Id into the data types directory.

4.3 Facts Directories

To continue the example from above: while the Data Type Directory contains the definition of items such as "currency number" the Facts Directories contains a list of 192 world currencies, and "facts" to describe what the currency transaction is about. The business app will choose the correct currency from the currencies directory, and it will choose appropriate further facts such as "sale" or "rent" to complete the description. The Facts Directories are intended to cover or model business (and human) activities and interests.

These directories will include date/time ranges for the validity of a fact. This could be for a extended period e.g. when the UK is or was a member of the EU and when not. Or the fact could change frequently as for exchange rates.

Pacio will develop and maintain the facts directories as an open source service accessible to all, maintained via a moderated folksonomy[9] like process, plus automated data feeds for changing data such as exchange rates

It is accepted that maintaining these directories will be subjected to some of the issues listed in the 2001 paper Metacrap: Putting the torch to seven straw-men of the meta-utopia mentioned in *The Semantic Web* section but by keeping them to "generally accepted facts" and moderating the folksonomy process it is hoped to minimise the issues or at least control them. Pacio will have a constitution and a Governance Council to guide its governance processes. It is expected that part of this will be devoted to the facts directory maintenance process.

The facts directories will be built to be reliable and available at all times using blockchain and distributed database technologies so that apps may depend upon them.

The facts directories will grow to become large, but in concept they will remain simple. They will be a major part of making SSIM easier to use and understand than other approaches which try to build relevant facts into domain specific ontologies or taxonomies.

Directories will use existing classifications where possible e.g. ISO country codes.

All directories will allow for groupings or classes within them e.g. Asian countries.

All facts will allow for language and jurisdictional variations.

Facts can be flexibly combined as a set of references and relationships which results in a single SID as described in the next section.

The vast majority of directory uses, certainly for business needs, will be for the simple description of an item of data using one or more references without need for the complicated semantic syntax of systems such as OWL (Web Ontology Language). An example is "this is an XCD money rent expense".

Relationships

The relationship for a data item being described with a "fact" will usually be self evident as adjectival (descriptive), or "is a", "is in", "has the property of", "is a member of", or "has as a member" according to the data type of the item and the type of the fact. Adjectival and "is a" are the commonest. For "this is an XCD money rent expense", the data type would be "money" with "XCD", "rent" and "expense" all being descriptive or having a "is a "relationship to the data item.

For "Saint Lucia is a country in the OECS (Organisation of Eastern Caribbean States)" the relationship of Saint Lucia to the OECS is understood to be "is a member of" because OECS would be defined as a fact that is a group or collection of countries. Whereas "OECS" plus the fact "Saint Lucia" are linked by a "has as a member" relationship.

Such simple self-evident relationships, or relationships derived from the data type of the item and the type of a fact, meet business reporting needs, if not all those of other more esoteric domains such as metaphysics, pure mathematics, or religion. SSIM development will start with just these simple relationships.

Support for other relationships such as "is not", "is not a member of set/class x", plus logical combinations using "and" or "or" could be added as requirements arise or are defined. The reason for deferring such additions is that some other semantic and data description languages have suffered from trying to define everything at the start, with the result that options became over complicated and have never been used in practice, resulting in later specifications advising against their use, and their ultimate deprecation. With SSIM the approach will be to add more complicated options only if there is proven to be a real need for them.

Business Related Directories

The initial directories to be developed are those needed for business and financial reporting, though some of these will also be applicable to other needs:

- Types of directory fact e.g. country, state/province/ political grouping of countries, set of x, etc for each of the following directories
- Countries, states/provinces, jurisdictions, regions, cities/towns, and groupings of these
- Human languages
- FIAT currencies
- Crypto currencies and their blockchains or other distributed ledger systems
- Functional roles – expense, sale, equity, fixed asset etc
- People's roles and occupations - director, partner, officer, remote worker etc
- Entity types from sole proprietorship to public limited company including charities, NGOs, and Government Departments/agencies, for all the variations in the world
- Stock and futures exchanges where public companies may be listed
- Crypto exchanges
- Industry/business classifications
- Units of weight and measure
- Activities – all business and human activities e.g. 'retail sale', 'accounting', programming, exercising etc, keyed to units where applicable e.g. truck driving and tonne kilometres
- Product categories e.g. household appliances > climate control appliances > fans > ceiling fans
- Product barcodes including UPC (Universal Product Code), EAN (International Article Number, previously European Article Number), ISBN (International Standard Book Number), Code 128 etc. Of the order of 200 million such codes exist with new ones being created constantly. They may be accessed via online APIs for use by SSIM.
- Service types – accounting, auditing, legal work etc
- Facts and terms relevant to so called ESG (Environmental Social Governance) reporting
- Report names or headings such as "Balance Sheet"

General Data Directories

More general directories not so specific to business data will also be developed progressively, covering such things as:

- Colours
- Materials
- Terms for the sciences: physics, chemistry, biology, botany, geology, genetics, mathematics, astronomy, palaeontology, ecology, oceanography, meteorology, zoology etc
- Medical terms
- Religions
- Product names

- People names
- Dynamic facts from data feeds e.g. exchange rates, stock prices, crypto prices …
- Imported facts from knowledge graphs and public datasets such as those listed in the *Knowledge Graphs and Public Datasets* section
- Other groupings of facts that people may be interested in and are prepared to help build

4.4 SSIM Id or SID

Any item of data is classified or semantically described using a single 64 bit (8 byte) number called an SSIM Id or SID. SIDs allow totally flexible data description in just 8 bytes, which will help make data storage and processing efficient. No long or variable length tags are involved.

A SID is a key into a global database of sets of references or keys into the Facts Directories, and optionally for data or an app using it, an SSIM Ontology reference as described in section ??.

Each set of references used by an app results in a single SID. A SID provides context and semantic content information.

A SID provides the equivalent of XBRL context plus concept tags and attributes.

64 bit SIDs allows for 2^{64} - 1 or 18,446,744,073,709,551,615 different sets of references which is 1.8 Billion for every person on earth at projected peak population of 10 Billion people. That should cover the world's needs for decades. If ever that limit should be approached, then extending SIDs to just 10 or so bytes would allow for centuries of reference combinations.

SIDs once used will exist "forever". They may become deprecated or no longer valid for new data, but will be kept indefinitely for historical analysis purposes.

The SID database will be a critical component of SSIM. Additions and updates will be logged for security and to enabling rebuilding in the event of catastrophic loss of the database.

A SID by itself would not tell a human observer anything, but software will easily show its references via the Pacio SID Service. Applications will be able to search or query by facts using indices.

SIDs could be shown as QR codes if an application wished to publish them.

SIDs are not specific to Pacio i.e. they can also be used by non-Pacio systems.

5 Pacio Transactions with SSIM

A transaction is the basic or lowest level Pacio record or data item. Everything starts as a transaction. Pacio is transaction driven, and SSIM applies at that level

A transaction may also be all that is involved. For a crypto transfer, or one piece of information, the transaction forms the complete record. For more complex storage needs discussed in the next section, the transaction also updates other storage elements.

A transaction includes:

- An Id – the transaction Id
- A date and time, called a datetime. All Pacio and SSIM datetimes use GMT/UTC time. Datetimes may be converted to local or other times for reporting purposes, but internally to keep things clear, GMT/UTC rules.

- Digital Ids or digids (section 4.1) for the entity (if there is one involved), the user/account, and the app/dapp which created the transaction
- Digids for who or what the transaction is going to if different from the creator, including for transactions with other blockchains
- A data type Id
- One or more binary data fields according to the transaction's data type
- A SID to describe the transaction. The SID is generated by the app which creates the transaction. For transaction only storage the SID involves just references from the facts directories as described in SSIM Basics above. For transactions which update another storage element (next sections), a delement, the SID also includes an ontology (section **Error! Reference source not f ound.**) reference.
- A delement Id for the data set (section **Error! Reference source not found.**) if one is involved

Transactions will be retrievable by Id, and by indices for digids, SIDs, and for the references making up a SID. These indices will be large but stored totally off chain as they are non-critical and could be rebuilt if necessary.

Applications for FIAT or crypto transfers, and applications involving just storage of discrete items of knowledge will need northing more than transactions with SSIM. Business and other applications which involve more complex data storage and reporting, will, however, need to make use of the storage features described in the following section.

6 Pacio Application Data with SSIM

As an example of an application requiring more complex data storage and reporting is the production of company financial statements. That involves more than just transactions or single items of data. It involves aggregated data from numerous sources (today an average of over 800 spreadsheets in a Fortune 500 company) structured according to a set of rules – the accounting standard being followed, typically as defined via an XBRL taxonomy. The result is then put into a presentation format so that it can be read by stakeholders and be sent to the authorities.

Applications of that magnitude will use some or all of the more advanced aspects of SSIM listed below. A business application involving raw transactions through to final financial statements or <IR> integrated reports, would use all of them.

The components are:

- Data elements
  A data element or delement is an item of data created or updated by transactions. It is stored independently from the transaction or transactions which create or update it. All delements could be recreated from their transactions, but delements provide a means of coping with the numbers and complications of the raw transactional data. One delement could represent from one to millions of transactions.

  A delement is like an account in a financial system, but as they are also used for non-financial data, the term 'data element' or 'delement' is used for them rather than 'account'.

  Delements with an ontology reference (below) inherit optional attributes from their ontology element for:
  - Static – cannot change after being created
  - Deprecatable – can be deprecated and then removed from service
  - Dynamic by replacement where a repeat transaction replaces the current value(s)
  - Dynamic by summation where a further transaction adds to the current value(s)

  - Double entry accounting delement meaning that a set of transactions involved in updating the delement must sum to zero, with the set also updating one or more other double entry delements
  - Triple entry accounting delement

- Data Sets
  Transactions for entities may be grouped into data sets where that is a requirement of the application.

  For example, accounting or financial reporting apps would use Data Sets for each set of balancing double entry money transactions, a set of journal entries, or one group of non-monetary postings.

  Another example is the set of line items making up a purchase order or an invoice, which itself would in turn be a member of a double entry data set.

  Applications which involve periods would use data sets by period.

- Ontologies
  SSIM uses ontologies as part of standardised semantic organising of data and for use by SSIM Smart Reporting Objects (Ssros) and SSIM Import Export Objects (Simeos) below.

  Ontology use is not mandatory – Pacio and SSIM as described to this point can be used without an Ontology. However, ontology use is required for the more advanced SSIM uses provided by Ssros and Simeos. Business applications which produce XBRL reports or work with other data description languages will need to use ontologies.

  SSIM ontologies are simpler than other ontology/taxonomy systems because of the information content delegated to other parts of SSIM, namely the Data Types Directory, Facts Directories, and Ssros plus Ssro Sets. Simply put, SSIM ontologies provide the framework, while Ssros and Ssro sets plus data types and facts fill in the details. Then Simeos provide interfaces to and from other systems.

  Any number of ontologies may be created. There will be many of them, ultimately thousands of them for all types of storage and reporting requirements. General purpose ontologies will be developed and maintained by Pacio in a moderated folksonomy process, but entities may also create specific purpose ontologies if they so wish.

  Ontologies may be created for any data organisational purpose, but the initial ones to be built by Pacio will be intended for business and financial data. These ontologies could be structured in accordance with an accounting standard e.g. a US GAAP focused one, or an IFRS focussed one.

  However, that is not necessary or even desirable given the capability of Simeos. It will, in fact, be a goal of Pacio to produce more general or fundamental ontologies without the anglo-american biases of the current IFRS and US GAAP views of world, and which can be used to generate reports according to any desired accounting standard. That is where the “standardised” part of SSIM comes into play.

  Ontologies use realms, domains, and elements to define the framework, with an element being the lowest level or most basic. An element corresponds to a Ssro or Ssros. There will typically be more Ssros than ontology elements, because Ssros can be categorised in other ways than just their ontology element, plus Ssros can be replicated as optionally filtered slaves for summing and presentations in various ways.

  Ontologies are maintained in spreadsheets which are imported into Pacio. No arcs as in XBRL or triples as in WC3 web 3 specifications and knowledge graphs are involved. The import performs validity checks to check for possible errors or inconsistencies. Spreadsheet use makes it easy for accountants and business people to visualise and understand how SSIM ontologies are constructed.

Ontologies provide the equivalent of the XBRL definition view. Presentation and calculation or other XBRL views are provided by Ssros and Ssro Sets in the SSIM case.

Apps will use ontologies to categorise the data they are creating, and, if alternatives should exist for a particular item, to prompt a user for a decision, all without users (other than ontology creators/maintainers) needing to know anything about the details of the ontology being used.

The relatively simple SSIM ontologies in conjunction with the other SSIM components will be more powerful than other methods of expressing semantic data relationships such as W3C's OWL ontologies or XBRL Taxonomies yet will be easier to work with and understand at the user level.

- SSIM Smart Reporting Objects or Ssros
  SSIM Smart Reporting Objects or Ssros pronounced “s-rows” and Ssro Sets are used for processing data for reporting or querying. They provide for intelligent, flexible reporting.

  The Pacio report writer uses them. An app’s own special purpose report writer could use them.

  Ssros do not store data – they are in memory objects for processing data from delements, and optionally transactions, for reporting and querying purposes.

  Ssros are organised according to an ontology, using realms, domains, and elements in a multi branch tree structure. Ssros embody knowledge about their environment derived from their ontology, facts, and attributes plus logic built into them by Ssro developers.

  Ssros may be copied to another place in the tree as a slave of the master Ssro, with optional filtering, including from transactions (at a speed penalty), to permit summing or reporting in flexible ways without having to create multiple data sets. For example, sales by region and sales by product/service type could be reported on from just the one set of sales data.

  Ssro Sets are sets or collections of Ssros to allow for library like modular reuse.

  Ssros and Ssro Sets are maintained in spreadsheets which are imported into Pacio. No arcs as in XBRL or triples as in WC3 web 3 specifications and knowledge graphs are involved. The import performs validity checks to check for possible errors or inconsistencies. Spreadsheet use makes it easy for accountants and business people to understand how Ssros and Ssro Sets are structured.

- SSIM Import Export Objects or Simeos
  SSIM Import Export Objects or Simeos pronounced “sim-e-oh-s” will allow export of data organised via one ontology to other systems, or the import of data from other systems.

  Conversion will only be possible to the extent that other systems can replicate the information depth of SSIM. In cases where the target system lacks equivalents for some SSIM features, information content will necessarily be lost on conversion.

  Simeos and Simeo Sets will work as extended Ssros and Ssro Sets that work with both the SSIM data and the target “ontology”.

  Writing or the creation of transactions, data sets (if applicable), and delements, will be optional if the target is an SSIM ontology. Otherwise the data will be held in the Simeos and Simeo Sets like Ssros, and be available for reporting from there, as for normal Ssros and Sssro sets.

  Use case examples are:

  - One SSIM ontology to another e.g. for conversion from the universal or fundamental SSIM ontology to a “US GAAP” or IFRS focussed one, with optional writing
  - Ontology upgrades e.g. from the 2018 ontology to the 2019 ontology, with writing expected unless testing
  - SSIM ontology to an XBRL taxonomy for XBRL reporting – no writing

  - SSIM ontology to W3C type ontologies for reporting – no writing
  - SSIM ontology to schemas as in Schema.org for reporting – no writing
  - SSIM ontology to any other community required "ontology" as per Ontologies Ontologies Everywhere – but Who Knows What to Think? for which a Simeo interface can be written
  - Importing data from a non SSIM based app to a Pacio SSIM based app using a SSIM ontology, with writing expected unless testing

  As for other SSIM components, Simeos and Simeo Sets are maintained in spreadsheets which are imported into Pacio, with extensive validity and consistency checking.

More details on these SSIM components are provided in a separate SSIM document.

7 SSIM Advantages vs the Alternatives

The digital world is going through a revolution with AI (artificial intelligence), IoT (the Internet of Things), distributed technologies (blockchain), almost universal internet availability via mobile phones, and faster speeds e.g. far faster 5G networks will launch in 2019, all of which will add to the digital data flood.

The current alternatives to SSIM described in the *Semantic Data Initiatives* section will struggle to cope, let alone help people gain the potential benefits that are possible. Continuing with them is akin to adding more stories to the Tower Of Pisa without fixing the foundations.

To cope and reap the potential rewards for mankind, a revolution in how data is categorised, stored, and accessed is needed.

The main advantage of SSIM vs the alternatives is that it is such a revolution, a green field fresh start, designed to scale to cope with the flood, while providing complete control and access, yet keeping things easy for users, and software developers too.

SSIM advantages which deliver this are:

- Using directories of facts as a key component of categorising data. These directories will become large and diverse, yet be simple to understand and use. None of the alternatives do this so directly and simply.
- Using distributed (blockchain) technology to make the directories of facts secure, reliable, and accessible to all. None of the alternatives do this.
- Incorporating complete flexibility from KISS approaches like JSON to as complex as an advanced app might want to make things, without being restricted by a base KISS level.
- Building in support for data types that the alternatives don't, or at least not in any standardised way which has been adopted. An example of this is activity data that is required for TARI® which can bring about 10%+ productivity improvements in businesses worldwide. (Productivity improvements are how wealth is created to lift people out of poverty and provide the means to tackle pressing world challenges such as climate change.)
- Recognising that the world will use thousand of currencies (counting cryptocurrencies and tokens of various kinds) in addition to the traditional FIAT currencies, and allowing for any mix of these, including reporting on them at any particular time vs any other currency, which is an especially big factor with volatile crypto currencies. The alternatives do not cover this at all.
- Recognising that AI and IoT use will result in huge numbers of transaction, eventually billions per second, where in total the knowledge in those transactions will be immense, but where any single one of them is pretty worthless, and being able to handle this. The alternatives will drown.

- Recognising that reporting requirements will become ever more demanding e.g. ESG (Environmental Social Governance) being added to financial statements, and allowing for this without the whole system becoming incomprehensible, even to those immersed in it.
- Keeping ontologies, where needed at all, simple enough to be viewed and maintained in spreadsheets, understandable to anyone who is comfortable with spreadsheets. Ontologies or taxonomies in the alternatives have become horrendously complicated, so much so as to be beyond the understanding of any one person, especially of a software engineer trying to develop an app to make sense of data!
- Building in the smarts via reporting objects which can be tweaked any way a particular entity or app might want, without having to rework things or break the bank with consultants
- Building in a similar object system to interface to any other data protocol in use, so providing ultimate flexibility, it being recognised that the world is a complex place, and no one standard, not even SSIM, is ever going to win them all!
- Being fully open, and staying flexible. The alternatives attempt to do this too, of course, but it is proposed that the structure of SSIM will make it easier to actually achieve the flexibility that a rapidly evolving digital world will need.

8 How SSIM will be used

SSIM will primarily be used by apps with apps defined broadly to mean software applications of all kinds, including mobile apps, blockchain dapps (distributed apps), web apps, and traditional business/computer applications.

Those apps will allow people to use data to do business, to learn, to play, to interact, and generally to make progress, with ownership/control of data returned to its owners/creators, and without fear, since the better security of SSIM (as implemented via the Pacio platform), will permit all this while reducing opportunities for the crooks of the world.

That IDC report[10] mentioned in *The Standardised Semantic Data Need* section predicted that “From 2018 to 2023, with new tools/platforms, more developers, agile methods, and lots of code reuse, 500 million new logical apps will be created, equal to the number built over the past 40 years.” Those 500 million apps are how SSIM will be used!

Those apps will cover every human activity. “There is an app for that” will be such a cliché that it would have dropped out of use. Apps, like phones, will just be there. And those apps will use SSIM to make data use easy and natural to the users of the apps. Certainly, app developers will need to understand SSIM, but that understanding will be easy.

It is impossible today to predict exactly what apps will be developed, and to what use SSIM data will be put. All that is certain is that there will be a myriad of ways, and thus the need for an extremely flexible, powerful, and secure system, yet one easy to use and understand.

[10] IDC Report 10 predictions for how the tech industry will change in 2019 and beyond

9 How SSIM will gain adoption

SSIM is just in its early stages of development. A prototype to demonstrate some aspects of SSIM is in development. 2019 is expected to see the design work completed and the first working code become available.

Development of the SSIM Directories of Facts which will be a major exercise. Estimates of the time and cost of building the directories is years and billions of dollars if done commercially[11] [12]. It can be expected that many people will volunteer to help, just as people have done for Wikipedia, and Freebase etc discussed in the *Knowledge Graphs and Public Datasets* section, which will help with the cost but not the time.

Then adoption will take time. A lot of inertia with current methodologies is involved, even if the people using them find them frustrating and know that there should be a better way.

Also, it is clear that as a proposal emanating from Pacio, adoption will depend to a large degree, at least initially, on the success of the Pacio platform. That will be influenced by factors other than SSIM i.e. how good the Pacio platform itself is, and the success it achieves in the market.

The Pacio goal is a distributed technology platform which will be attractive to app developers and end users alike, which will grow to hosting thousands and ultimately hundreds of thousands or millions of apps. Funds to be raised for Pacio will cover both development and marketing or ecosystem building to see this happen.

Given Pacio success, SSIM will gain use and recognition. Pacio open governance will help. As SSIM is developed and becomes stable, though evolving as requirements dictate, Pacio will propose its wider use via industry and standards bodies.

It is expected that a tipping point will be reached when the benefits become obvious, though how long that might take it is not yet possible to predict. It could happen within three years of Pacio launch targeted for late 2019/early 2020, or it could take a decade, given the inertia involved, despite ever increasing frustration and costs with the current alternatives.

Pacio is determined to see it through, however long it takes.

10 Conclusion

The world needs a better, revolutionary, system to manage and extract maximum benefit from the flood of data in the digital age, a system which categorises that data in a standardised semantic way, and which can scale to handle the volumes involved while maintaining security, and privacy.

The Standardised Semantic Information Model (SSIM) is proposed as that system.

Pacio Core Ltd
Saint Lucia, West Indies
https://www.pacio.io

[11] How much is a Triple? Estimating the Cost of Knowledge Graph Creation
[12] Constructing and mining web-scale knowledge graphs