CDISC 360i, and the WORM that Turned

Presented by
Jeremy Teoh, AD Data Asset Framework & Services, GSK
Warwick Benger, Director of Data Standards, GSK

Meet the Speakers

Warwick Benger

Title: Director, Data Standards

Organization: GSK

More than 25 years of experience in clinical data delivery and analysis, helping others to find their best self and putting people at the centre of data and technology.

Currently co-leads GSK's "end to end standards" transformation programme and metadata-driven pipeline

Shared: UK-based, Stat Programming, E2E Standards, Physics

Jeremy Teoh

Title: Associate Director, Data Asset Framework and Services

Organization: GSK

Aspiring ontologist, playing with CDISC metadata for over a decade

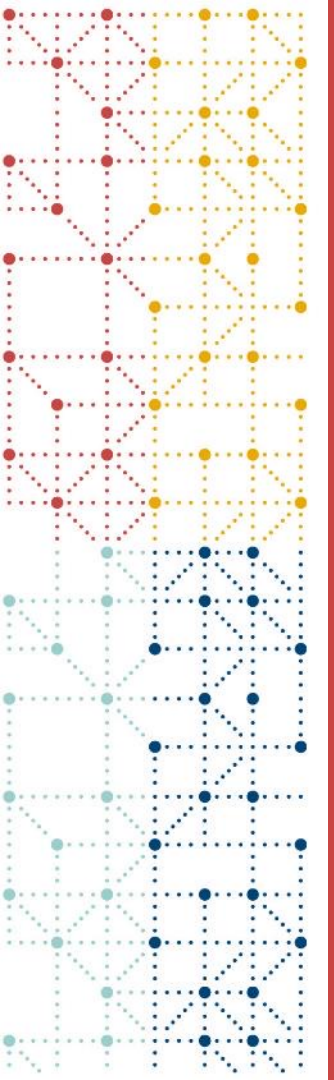
Modelling and strategy consultancy for CDISC Working Groups, Pharma, Biotech and Startups

CDISC 360i Run Team Co-lead

Disclaimer and Disclosures

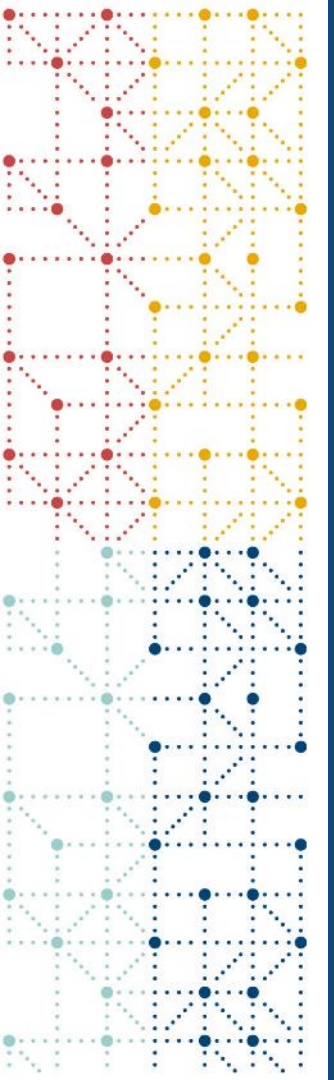
- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC or GSK.*

- *The author(s) have no real or apparent conflicts of interest to report.*



Agenda

1. Introduction
2. Unifying E2E Model
3. Walking the E2E Walk, Defining Analysis
4. Call to Action



Introduction: WORM

WORM

Write Once, Read Many

- “WORM” (Write Once, Read Many) has been a central tenet and ambition for the industry for a long time
- Fundamentally, once a thing (i.e. some data element) is defined, that definition can be reused through the life of that thing

Use Standards fully and they will bend

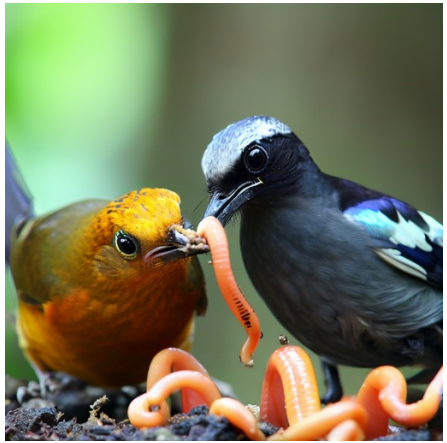
For more details see Appendix: Unifying Models by Challenging Assumptions

Extending ODM

- Using Biomedical Concepts to bridge gaps
  - USDM ←→ODM
  - odm:ItemData ←→ Value-level Definitions
  - Data Specialisations by linking ItemRef to BCProperty, ItemGroup to BC
- Adding name, label and semantics to logical statements
  - Reusable meaningful Conditions applied to USDM contexts such as BCs, Analysis Population definitions
- Supply WhereClause with reusable Conditions
  - ODMv2 has reusable pieces of logic resolving as Boolean
  - Construct logical statements from these pieces
    - multiple Conditions per WhereClause
    - multiple RangeChecks (linking specific Items and values) per Condition
    - support traceability of compound logical statements across USDM, Define, ARS

Extending USDM

- Codified the equivalence between *odm:Coding* and *usdm:Code*
- Reusing *odm:Condition* and *odm:CodeList*
  (are so many ways of representing CT needed?)
- *DetectedEntity* introduced for NER and missing links
  - Bridges unstructured text in USDM classes to
    - Code, Biomedical/Analysis/Derivation concept
    - Activity, Procedure, Scheduled Decision Instance
    - Administration
    - Indication, Condition
    - Value, Range
    - Population
  - Respects negation, i.e. “this thing is explicitly mentioned as **not** being relevant”
  - Bridges the missing links in USDM



How we used USDM to describe itself

Section Extraction from Protocol Documents

Extract custom sections (like objectives, endpoints, schedule of activities) from a protocol PDF, into a specific JSON format, ensure no facts are left behind. Robust and can handle complex SOA tables.

Table 3 Objectives and endpoints

Objectives	Endpoints
Primary	
Assess the effect of MMB in combination with LUSPA on TI response by Week 24 (Rolling TI response)	• Proportion of participants with TI by the end of Week 24; defined as not requiring RBC transfusion (except in the case of clinically overt bleeding) for any ≥12-week interval.
Secondary	
To characterize the safety and tolerability profile of MMB in combination with LUSPA	• Incidences of AEs, SAEs, and AEs leading to discontinuation. • Clinically important changes in laboratory parameters (hematology, chemistry), vital signs, and ECOG performance status.


```
"objectives": [
  {
    "text": "Assess the effect of MMB in combination with LUSPA on TI response by Week 24 (Rolling TI response).",
    "type": "Primary",
    "endpoints": [
      {
        "text": "Proportion of participants with TI by the end of Week 24; defined as not requiring RBC transfusion (except in the case of clinically overt bleeding) for any ≥12-week interval.",
        "type": "Primary"
      }
    ]
  },
```


Named Entity Recognition (NER) on Content

Perform NER on unstructured text in context of the identified USDM elements to enrich study with definitions, structure and relationships that would otherwise be missed

Subjects must have liver biopsy LAB TEST demonstrating NASH MEDICAL CONDITION with Brunt Stage 3 fibrosis MEDICAL CONDITION within 12 months of randomization.

The subject is >= 18 years DEMOGRAPHICS of age and <= 75 years DEMOGRAPHICS old at the time of screening.

The subject is willing and able to provide written informed consent.

The subject is not pregnant INTERVENTION STATE and must have a negative pregnancy test LAB TEST prior to start of the study.

Post-menopausal women DEMOGRAPHICS must have been amenorrheic MEDICAL CONDITION for at least 12 months to be considered of non-child-bearing potential.

```
Modifier at token 23 --> Entity at token 30
Modifier at token 30 --> Entity at token 58
Modifier at token 58 --> Entity at token 64
Modifier at token 64 --> Entity at token 75
```

How we used USDM to describe itself

Target content and context described according to CDISC resources

USDM Structure description (LinkML)

USDM IG (PDF)

Controlled Terminology lookups included (LinkML)

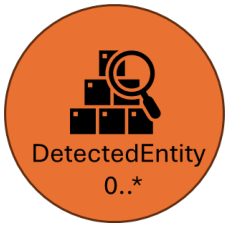
Full definitions included from references (LinkML, UMLS lookup, CT)

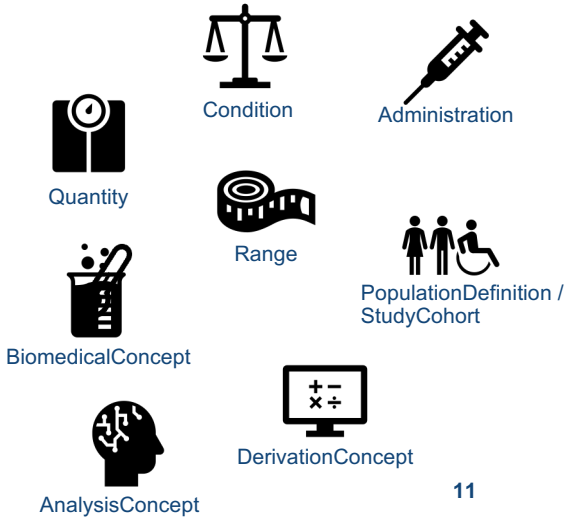
Missing links patched into USDM

USDM can describe reusable components and qualifiers that get referenced in unstructured text

The **DetectedEntity** class extends USDM to bridge meaningful relationships between study design components, that are missing from the current model / implementation

Enrich structured protocol context & content … by detected components in unstructured text… that contain recognisable structured content.





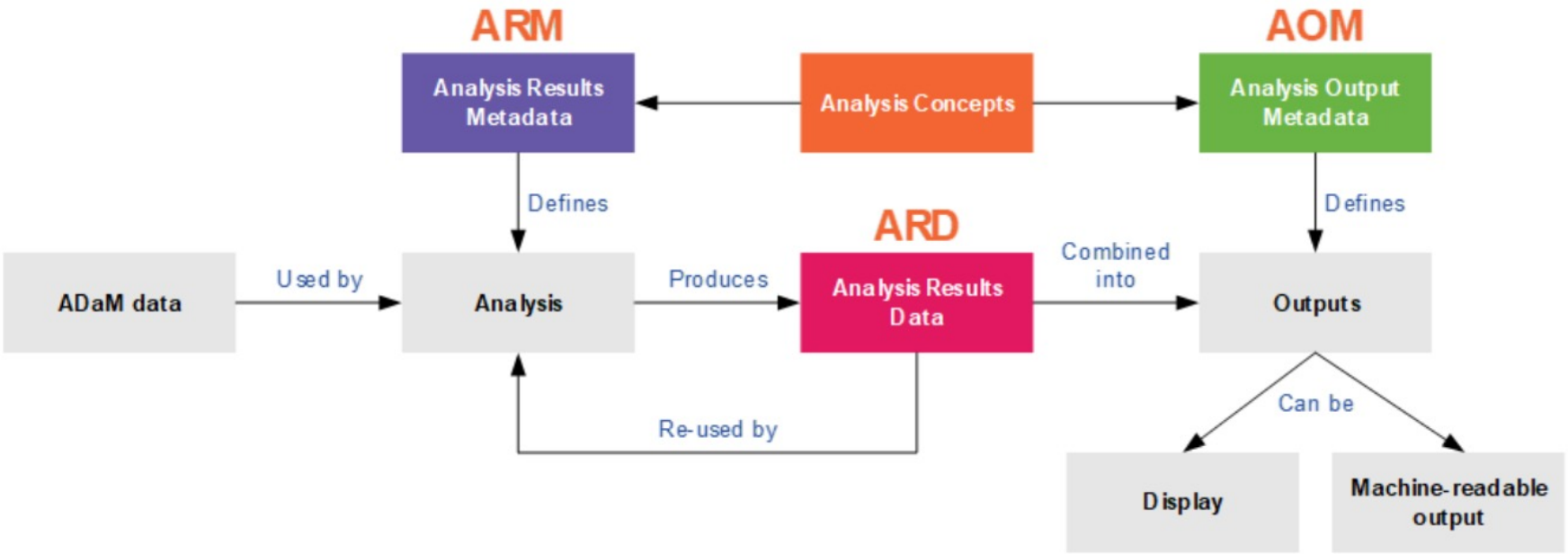
A Proactive, Participatory approach is needed

The early bird catches the worm

- Put a metadata-first approach into everything
- Define and share content as far upstream as possible
- It's not reasonable to expect SDOs like CDISC to be on the critical path for new concepts

ARMADA Model

Analysis Results Metadata And Data Automation



Walking the End-to-End Walk

Robust Governance & Digital Capture	Implementing an E2E philosophy requires strong governance and comprehensive digital capture of study definitions, ensuring data flows and submission content are automated from study design to results at GSK.
Stakeholder Engagement	Harmonized data practices are facilitated through collaboration among Clinical, Data Management, and Regulatory domains, enabling reuse and interoperability.
Role of CodeList & Standards	The setup of CodeList standards may be seen either as a programming activity or part of the clinical scientist's role, involving ClinOps.
Digital Collection & Publication	Collect all analyses digitally, register, and publish biomedical and analysis concepts to facilitate robust data management.
Key Areas & Management Systems	Engage CTO, MW (smart protocol), Standards, Coding, Master Data Management, Ontology Management, Data Quality, and Data Fabric.
Harmonized Data Access	Develop standardized methods for accessing data.
Mapping Descriptions	Create generalized mapping descriptions applicable to various data sets, beyond just SDTM / ADaM.
Terminology Standards	Move towards a unified collection and curation of terminology standards.
Collaboration & Contribution	Integrate RWE, DM, CP, contribute BCs to COSMoS, collaborate on shared libraries, and share analysis concepts throughout the industry.

Call to Action

- The CDISC worm is turning
- Moving towards expressing things through data models and unique reusable concepts, distinct from data implementations (specialisations)
- Industry can help by committing to and aligning to this transition, and adopting the underlying models

Thank You!

Appendix:
Unifying Models by Challenging Assumptions

Are CDISC's models really interoperable yet?

Have we been using existing standards to their full potential?

Assumption 1: ODM is only for CRFs

ODM is the base model for Define-XML, used in SEND, SDTM, ADaM

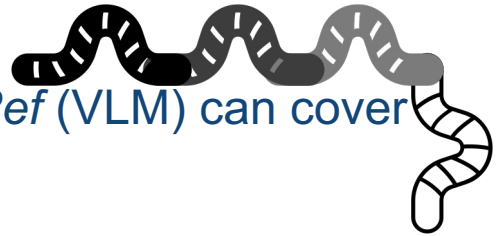
What better model for a specification than the one used for submission?
That way consistency, referential integrity is guaranteed E2E

In general, ODM serves as an ideal basis for specs and data contracts

- ODMv1 can describe any tabular dataset metadata
- ODMv2 can represent nested dataset metadata (e.g. BCs), with traceability

Assumption 2: VLM is only for xxORRES-type variables

The decision back in the Define 1.0 template to limit VLM to a narrow set of examples has led to an assumption that Value-Level Definitions are only for those variables



Remove this assumption, and *odm:ItemRef* (VLM) can cover

- Any field+row of any dataset*
- Data Specialisations of BCs
- Domain-wide variable definitions**
- Data Contracts

*even nested ones with v2+
**strictly-speaking, an Item in a Domain context is an ItemRef

The importance of decoupling

Representation != Meaning != Context

Supply != Demand

Condition != WhereClause

Confusing these things prevents WORM (Write Once Read Many)

The same entity can mean different things to different contexts

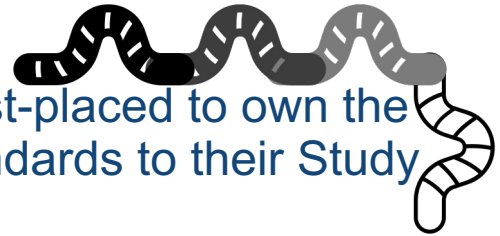
Assumption 3: Standards define content and meaning

When applied to a real context, the language and set of assumptions attached to that context colour the true meaning of what is represented

Standards define core *expectations, structure* and *rules*
Templates and prior studies describe reusable content from different contexts

Content and meaning are *context-dependent*

Study teams know their own context; they are best-placed to own the nuanced application of content, meaning and standards to their Study



Supply != Demand

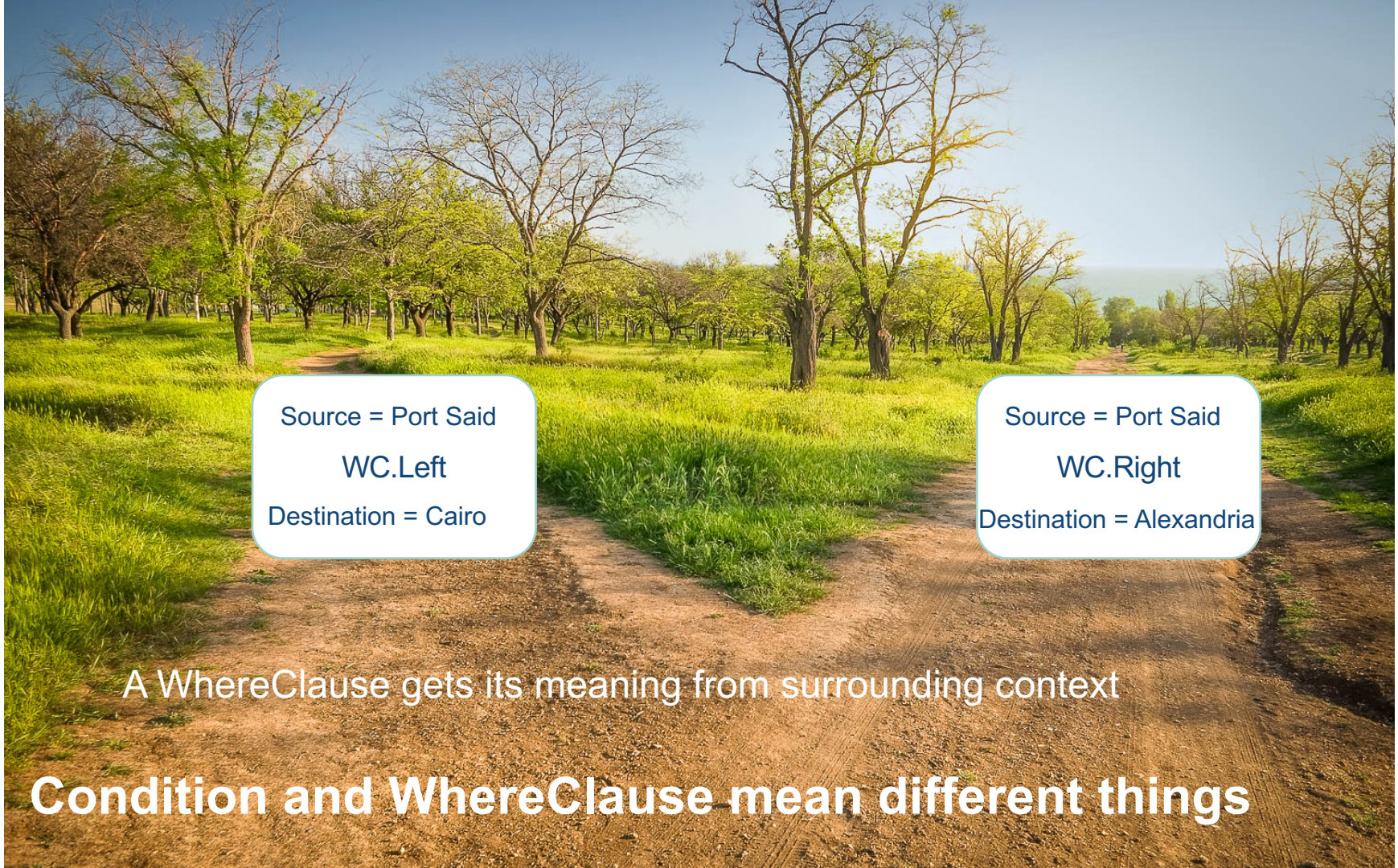
Need for a TFL or Dataset can be described prior to the existence of the data and metadata assets that supply it

A data context demands a logical data definition (Where Clause, use OR if >1)
that is satisfied by one of more logical statements (Condition / RangeCheck, use AND if >1)

Separating the right things can increase connectivity

Independent meaning allows disparate representations of the same thing to be linked to the common meaning

When meaning of statements is independent from that of their context, the statements become portable i.e. meaning can be tagged directly and travel with them



A WhereClause gets its meaning from surrounding context

Condition and WhereClause mean different things

Using ODM + BCs for Granular Data Definition

- What coordinate system pinpoints data from any dataset?
  - “X-axis”: Column name
  - “Y-axis”: Logical condition describing row (Time, Context)
  - Non-tabular bonus: Nested objects with OIDs
- What is this data point?
  - What does it mean? – Concept*, ConceptProperty*, Data Element Concept*
  - How is it implemented? – ItemGroup (Context), ItemRef (Contextual Definition), Item (Column name)
- How is this specific data point derived?
  - Current odm:ItemData references Item (column) rather than ItemRef (implementation of BC Property as VLM) - **need to introduce BC-level ItemGroup context as the missing link**
  - A WhereClause linked to BC contextualises that data point against specific VLM
- What logical condition describes this context?
  - WhereClause is a distinct implementation that satisfies a *demand*, an OR contribution
- What logical condition connects data to multiple usage contexts?
  - Both Condition and RangeCheck are Boolean logical structure that may reference further structure, an AND contribution that *supplies* multiple WhereClause contexts

*applicable to Biomedical, Analysis, and Derivation Concept dimensions

Assumption 4: Standards are for staying within

Fully-understanding a standard requires testing its limits and figuring out how to overcome them

Standards need interpreting for each new context.

Not all contexts are known up-front, standard content can never be complete

<u>Effort spent repeating established work is a waste of time:</u>

CDISC implementation effort needs to be less about learning what about has gone before, and more about defining what is new well-enough for it to be (F)indable, (A)ccessible, (I)nteroperable, and (R)eusable.