

SchemaCom – An XML Schema Comparator

Ihe Onwuka

Systems Engineer
LS Technologies

Ihe.Onwuka@gmail.com

About Me

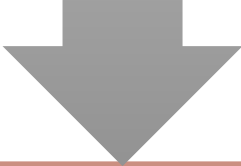
Been working with XML since 2003.

Was tricked into learning about XML by our then Consulting Director – who told me “It’s just like COBOL”


Currently System Engineer with LS Technologies architecting complex data models for the US Federal government

SchemaCom What's Your Problem Then

Schemas change but how they've changed isn't always communicated



Need to port existing data to be conformant with the new version of the schema.



Need to know how and what to map between instances of schema versions

Schema Comparison is a hard problem

That's quite a challenge, because so many differences are possible: you may have to make compromises depending on what you actually encounter. (For example, you may or may not need to bother with named model groups, depending on whether either of the schemas you are comparing actually uses them. Most schemas don't.)

<https://softwarerecs.stackexchange.com/questions/48067/diff-tool-for-xml-schema-file>


You really don't want to do this by hand and for sure nobody is going to write such a complex method from scratch, it doesn't fit the [Q&A nature of StackOverflow](#). However, several tools exist that can do this for you:

<https://stackoverflow.com/questions/32522975/comparing-two-xsd-files-for-differences-in-elements>

<https://docs.deltaxml.com/xml-compare/latest/comparing-xml-schema-15008896.html>

Schema Comparison – Survey of whats out there

- schemaDiff - <https://www.membrane-soa.org/soa-model-doc/1.4/cmd-tool/schemadiff-tool.htm>

 The changes break the interface.

⊖ Element createResponse has changed:

⊖ ComplexType has changed:

⊖ Sequence has changed:

⚡⊕ Element NewElementForTest added.

⊖ ComplexType GetAllType has changed:

✓⊖ Annotation removed.

⊖ Imported schema <http://predic8.com/material/1/> has changed:

⊕ Element manufacturer added.

⊖ ComplexType ArticleType has changed:

✓⊕ Annotation added.

⊖ Sequence has changed:

⚡ Position of element name changed from 1 to 2.

⚡ Position of element description changed from 2 to 3.

⚡ Position of element price changed from 3 to 5.

⚡ Position of element id changed from 4 to 1.

⚡⊕ Element availability added.

⊖ Imported schema <http://predic8.com/common/1/> has changed:

⊖ ComplexType PersonType has changed:

⊖ Sequence has changed:

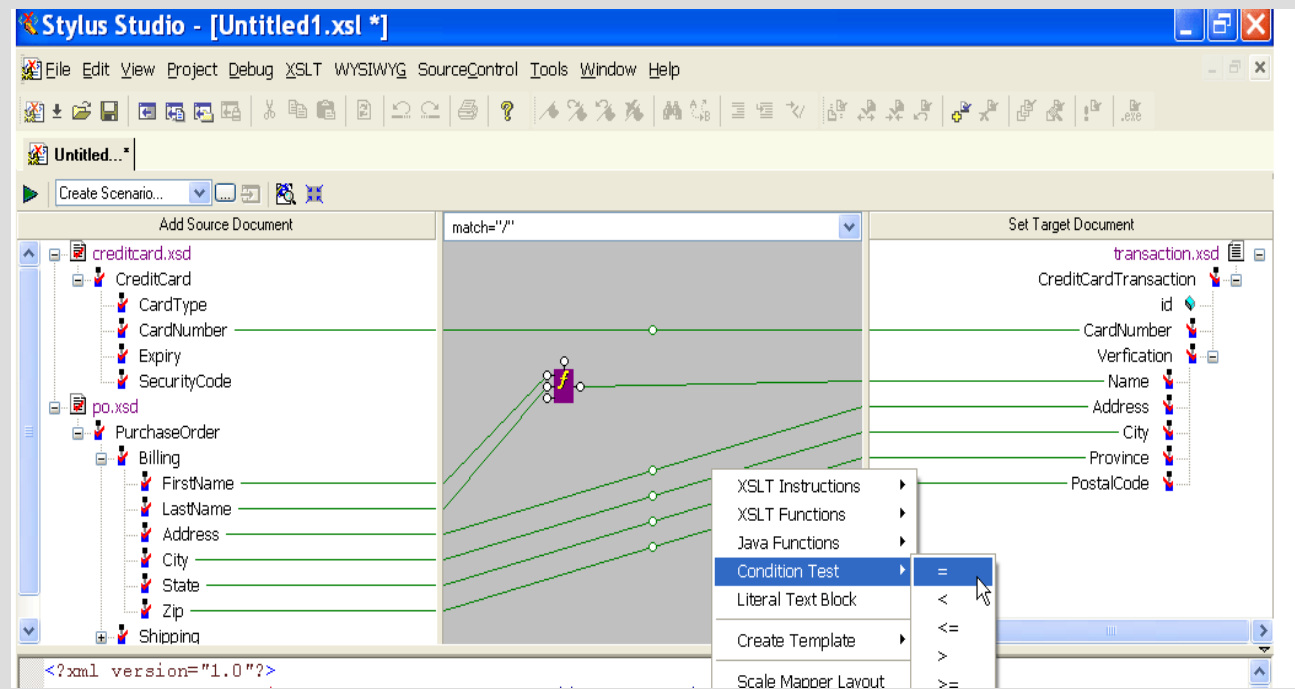
⊖ Element id has changed:

⚡ The type of element 'id' has changed from xsd:string to tns:IdentifierType.

Schema Comparison – Survey of whats out there

- Stylus Studio –
<http://www.stylusstudio.com/xsd-mapping.html>

- https://www.altova.com/diffdog/XML_Schema_compare



What's missing

No guidance on how to populate what the tools tell us has been added.

Several ways in which schemas can change without impacting mapping.

- If an annotation is added or altered does it affect the mapping?
- If the ordering in a content model changes does it affect the mapping?
- If the type of primitive changes does it affect the mapping?

The information that these tools give us on changes is limited

Tools that allow us to do the mappings but offer no guidance on what the mapping should be.



Mapping Aids

What does FeatureType map to in the new version

Select Schemas to Compare Search Types To expand a types details click on it's button below

Source: Fixm/3.0-Target: Fixm/4.0

Dropped ComplexTypes

FeatureType

FeatureMessageType

New ComplexTypes

Matched ComplexTypes

Identical ComplexTypes

Mapping Aids

What does FeatureType map to in the new version

Select Schemas to Compare fixm/3.0 vs fixm/4.0 Search Types fea To expand a types details click on it's button below

Source: Fixm/3.0-Target: Fixm/4.0

Dropped Complex Types

New Complex Types

Matched Complex Types

Identical Complex Types

FeatureType

@centre(fb:ProvenanceCentre Type)

Provenance Type(complex Type)

@source(fb:ProvenanceSource Type)

NasInBlockEstimated Type(complex Type)

NasRunwayArrivalEstimated Type(complex Type)

NasRunwayActual Type(complex Type)

NasOffBlockEstimated Type(complex Type)

NasRunwayDepartureEstimated Type(complex Type)

Provenance Type(complex Type)

@system(fb:ProvenanceSystem Type)

Provenance Type(complex Type)

@timestamp(ff: Time Type)

Provenance Type(complex Type)

Trigger Type(complex Type)

FeatureMessageType

A Mapping Aid

Select Schemas to Compare

Search Types

To expand a types details click on it's button below

Source: Fixm/3.0-Target: Fixm/4.0

Dropped Complex Types

FeatureType

TimeSequenceType

New Complex Types

ProvenanceType

Matched Complex Types

MessageMetadataType-fixm/4.0

Identical Complex Types

FeatureType Became ProvenanceType Who knew?

Select Schemas to Compare fixm/3.0 vs fixm/4.0 Search Types prov To expand a types details click on it's button below

Source: Fixm/3.0-Target: Fixm/4.0

Dropped Complex Types

FeatureType

TimeSequenceType

New Complex Types

ProvenanceType

Matched Complex Types

MessageMetadataType-fixm/4.0

Identical Complex Types

Select Schemas to Compare fixm/3.0 vs fixm/4.0 Search Types prov To expand a types details click on it's button below

Source: Fixm/3.0-Target: Fixm/4.0

Dropped Complex Types

FeatureType

@centre(fb:ProvenanceCentre Type)

ProvenanceType(complexType)

@source(fb:ProvenanceSource Type)

NasInBlockEstimatedType(complexType)

NasRunwayArrivalEstimatedType(complexType)

NasRunwayActualType(complexType)

NasOffBlockEstimatedType(complexType)

NasRunwayDepartureEstimatedType(complexType)

ProvenanceType(complexType)

@system(fb:ProvenanceSystem Type)

ProvenanceType(complexType)

@timestamp(ff: Time Type)

ProvenanceType(complexType)

TriggerType(complexType)

TimeSequenceType

New Complex Types

ProvenanceType

@centre(nas:ProvenanceCentre Type)

FeatureType(complexType)

@source(nas:ProvenanceSource Type)

FeatureType(complexType)

@system(nas:ProvenanceSystem Type)

FeatureType(complexType)

timestamp(fb: Time Type)

FeatureType(complexType)

Matched Complex Types

MessageMetadataType-fixm/4.0

Identical Complex Types

Mappings that are hard to spot

Select Schemas to Compare Search Types To expand a types details click on it's button below

Source: Fixm/4.0-Target: Fixm/4.1

Dropped ComplexTypes

New ComplexTypes

-
-

Matched ComplexTypes

stddsSpotIn(fb: Time Type)
SurfaceEventsType(complex Type)

@stddsSpotIn(fb: Time Type)
SurfaceEventsType(complex Type)

stddsSpotOut(fb: Time Type)
SurfaceEventsType(complex Type)

stddsWheelsOff(fb: Time Type)
SurfaceEventsType(complex Type)

stddsWheelsOn(fb: Time Type)
SurfaceEventsType(complex Type)

@stddsSpotOut(fb: Time Type)
SurfaceEventsType(complex Type)

@stddsWheelsOff(fb: Time Type)
SurfaceEventsType(complex Type)

@stddsWheelsOn(fb: Time Type)
SurfaceEventsType(complex Type)

Identical ComplexTypes

Mappings that are hard to spot

Select Schemas to Compare fixm/4.0 vs fixm/4.1 Search Types datal To expand a types details click on it's button below

Source: Fixm/4.0-Target: Fixm/4.1

Dropped ComplexTypes

New ComplexTypes

Matched ComplexTypes

Identical ComplexTypes

CommunicationCapabilitiesType-fixm/4.0
@otherDataLinkCapabilities(fb:CharacterString Type)
CommunicationCapabilitiesType-fixm/4.1
@otherDatalinkCapabilities(fb:CharacterString Type)

Select Schemas to Compare fixm/3.0 vs fixm/4.0 Search Types fac To expand a types details click on it's button below

Source: Fixm/3.0-Target: Fixm/4.0

Dropped ComplexTypes

New ComplexTypes

Matched ComplexTypes

Identical ComplexTypes

SurfaceEventsType

TelephoneContactType-fixm/3.0
@facimile(ff:TextPhone Type)
TelephoneContactType-fixm/4.0
@facsimile(fb:TextPhone Type)
NasFlightType-fixm/4.0

Superset content models

Select Schemas to Compare **fixm/4.0 vs fixm/4.1** Search Types To expand a types details click on it's button below

Source: Fixm/4.0-Target: Fixm/4.1

Dropped ComplexTypes	New ComplexTypes	Matched ComplexTypes	Identical ComplexTypes
<ul style="list-style-type: none">VerticalDistanceTypeTimeTypeOperationalAcceptabilityTypeNasDestinationTypeNasInBlockEstimatedTypeNasRunwayActualTypeNasOffBlockEstimatedType	<ul style="list-style-type: none">ExtensibleTypeSupplementaryDataSourceTypeSidStarAbbreviatedDesignatorToNextElementTypeFilingStatusTypeNasRunwayInformationTypeNasStandInformationTypeReportedTimeType	<ul style="list-style-type: none">AerodromeReferenceType-fixm/4.1OtherReferenceType-fixm/4.0DesignatedPointOrNavaidType-fixm/4.1RelativePointType-fixm/4.1HeightType-fixm/4.1TemporalRangeType-fixm/4.0TemporalRangeType-fixm/4.1	<ul style="list-style-type: none">ContactInformationTypeOnlineContactTypePostalAddressTypeTelephoneContactTypeIcaoAerodromeReferenceTypeGeographicalPositionTypePositionPointTypeSignificantPointTypeExtensionTypeAltitudeTypeAngleTypeBearingType

A Validation Aid – Beware false positives caused by namespaces

Select Schemas to Compare Search Types To expand a types details click on it's button below

Source: Saml/1.1-Target: Saml/2.0

Dropped Complex Types	New Complex Types	Matched Complex Types	Identical Complex Types
<div>AudienceRestrictionConditionType</div> <div>DoNotCacheConditionType</div> <div>SubjectStatementAbstractType</div> <div>NameIdentifierType</div> <div>AuthenticationStatementType</div> <div>AuthorityBindingType</div> <div>AuthorizationDecisionStatementType</div> <div>AttributeDesignatorType</div> <div>AuthenticationQueryType</div> <div>AuthorizationDecisionQueryType</div> <div>ResponseAbstractType</div>	<div>BaseIDAbstractType</div> <div>NameIDType</div> <div>SubjectConfirmationDataType</div> <div>KeyInfoConfirmationDataType</div> <div>AudienceRestrictionType</div> <div>OneTimeUseType</div> <div>ProxyRestrictionType</div> <div>AuthnStatementType</div> <div>AuthnContextType</div> <div>AuthzDecisionStatementType</div> <div>SharedSecretChallengeResponseType</div> <div>PublicKeyType</div> <div>AuthnContextDeclarationBaseType</div> <div>IdentificationType</div> <div>TechnicalProtectionBaseType</div> <div>OperationalProtectionType</div> <div>AuthnMethodBaseType</div> <div>GoverningAgreementsType</div> <div>GoverningAgreementRefType</div> <div>PrincipalAuthenticationMechanismType</div> <div>AuthenticatorBaseType</div> <div>ComplexAuthenticatorType</div> <div>AuthenticatorTransportProtocolType</div> <div>KeyActivationType</div> <div>KeySharingType</div> <div>PrivateKeyProtectionType</div> <div>PasswordType</div> <div>RestrictedPasswordType</div>	<div>AssertionType-saml/1.1</div> <div>AssertionType-saml/2.0</div> <div>ConditionsType-saml/1.1</div> <div>ConditionsType-saml/2.0</div> <div>AdviceType-saml/1.1</div> <div>AdviceType-saml/2.0</div> <div>SubjectType-saml/1.1</div> <div>SubjectType-saml/2.0</div> <div>SubjectConfirmationType-saml/1.1</div> <div>SubjectConfirmationType-saml/2.0</div> <div>SubjectLocalityType-saml/1.1</div> <div>SubjectLocalityType-saml/2.0</div> <div>EvidenceType-saml/1.1</div> <div>EvidenceType-saml/2.0</div> <div>AttributeStatementType-saml/1.1</div> <div>AttributeStatementType-saml/2.0</div> <div>AttributeType-saml/1.1</div> <div>AttributeType-saml/2.0</div> <div>RequestAbstractType-saml/1.1</div> <div>RequestAbstractType-saml/2.0</div> <div>RequestType-saml/1.1</div> <div>RequestType-saml/2.0</div> <div>SubjectQueryAbstractType-saml/1.1</div> <div>SubjectQueryAbstractType-saml/2.0</div> <div>AttributeQueryType-saml/1.1</div> <div>ResponseType-saml/1.1</div> <div>ResponseType-saml/2.0</div> <div>ResponseType-saml/2.0</div>	<div>ConditionAbstractType</div> <div>StatementAbstractType</div> <div>ActionType</div> <div>StatusType</div> <div>StatusDetailType</div>

Is that a duplicate definition of ResponseType or a false positive because namespaces were ignored.

Project Management Uses – Identifying Churn Rates

Select Schemas to Compare Search Types To expand a types details click on it's button below

Source: Fixm/4.0-Target: Fixm/4.1

Dropped Complex Types	New Complex Types	Matched Complex Types	Identical Complex Types
VerticalDistanceType	ExtensibleType	AerodromeReferenceType-fixm/4.1	ContactInformationType
TimeType	SupplementaryDataSourceType	OtherReferenceType-fixm/4.0	OnlineContactType
OperationalAcceptabilityType	SidStarAbbreviatedDesignatorToNextElementType	DesignatedPointOrNavaidType-fixm/4.1	PostalAddressType
NasDestinationType	FilingStatusType	RelativePointType-fixm/4.1	TelephoneContactType
NasInBlockEstimatedType	NasRunwayInformationType	HeightType-fixm/4.1	IcaoAerodromeReferenceType
NasRunwayActualType	NasStandInformationType	TemporalRangeType-fixm/4.0	GeographicalPositionType
NasOffBlockEstimatedType	ReportedTimeType	TemporalRangeType-fixm/4.1	PositionPointType
ArrivalMovementAreaHoldInformationType	DepartureDelayType	AircraftType-fixm/4.0	SignificantPointType
DepartureMovementAreaHoldInformationType	MovementAreaHoldInformationType	AircraftType-fixm/4.1	ExtensionType
FlightIntentType	AccelerationType	ArrivalType-fixm/4.0	AltitudeType
	NasIndicatedAirspeedType	ArrivalType-fixm/4.1	AngleType
	NasVerticalRateType	CommunicationCapabilitiesType-fixm/4.0	BearingType
	AsdexConfidenceType	CommunicationCapabilitiesType-fixm/4.1	DimensionsType
	MessageProvenanceType	DinghiesType-fixm/4.1	DistanceType
	NasAccelerationType	DepartureType-fixm/4.0	FlightLevelType
	TfdmAtcFlightStateType	DepartureType-fixm/4.1	FrequencyType
	TfdmFlightStateType	LastContactType-fixm/4.0	GroundSpeedType
	ArrivalTaxiOperationsMetricsType	LastContactType-fixm/4.1	IndicatedAirspeedType
	DepartureTaxiOperationsMetricsType	LastPositionReportType-fixm/4.0	LengthType
	TfdmArrivalType	LastPositionReportType-fixm/4.1	PressureType
	TfdmDepartureType	BoundaryCrossingType-fixm/4.0	RadioactivityType
	TfdmFlightType	BoundaryCrossingType-fixm/4.1	TemperatureType
		EnRouteType-fixm/4.1	TrueAirspeedType
		FlightType-fixm/4.1	VerticalRateType
		SupplementaryDataType-fixm/4.1	VolumeType
		AbstractConstraintType-fixm/4.1	WeightType
		AbstractRouteChangeType-fixm/4.1	WindDirectionType
		EstimatedElapsedTimeType-fixm/4.1	WindSpeedType
			AircraftOperatorType
			AtcUnitReferenceType
			IcaoUnitReferenceType
			OtherUnitReferenceType
			PersonOrOrganizationType
			ColourChoiceType
			FlightLevelOrAltitudeType
			FlightLevelOrAltitudeChoiceType
			TemporalChoiceType
			TrueAirspeedChoiceType
			TrueAirspeedRangeType
			VerticalRangeType
			GloballyUniqueFlightIdentifierType
			SsrCodeType
			AircraftTypeType
			AircraftTypeReferenceType
			IcaoAircraftTypeReferenceType

Above shows less volatility between v4.0 and 4.1

Project Management Uses – Identifying Churn Rates

Select Schemas to Compare: fixm/3.0 vs fixm/4.3 Search Types To expand a types details click on it's button below

Source: Fixm/3.0-Target: Fixm/4.3

Dropped Complex Types	New Complex Types	Matched Complex Types	Identical Complex Types
IcaoAerodromeReferenceType	MessageType	AerodromeReferenceType-fixm/4.3	NameValueListType
RunwayPositionAndTimeType	NetworkChoiceType	AtcUnitReferenceType-fixm/3.0	PointoutType
StandPositionAndTimeType	AirspaceDesignatorType	AtcUnitReferenceType-fixm/4.3	
UnlistedAerodromeReferenceType	DesignatedPointType	RelativePointType-fixm/3.0	
IdentifiedUnitReferenceType	GeographicalPositionType	RelativePointType-fixm/4.3	
UnknownUnitReferenceType	NavaidType	ReportedTimeType-fixm/3.0	
ExtensionType	RouteDesignatorType	ReportedTimeType-fixm/4.3	
FeatureType	RunwayDirectionDesignatorType	NameValuePairType-fixm/3.0	
DirectionType	SidStarReferenceType	NameValuePairType-fixm/4.3	
FixPointType	SignificantPointChoiceType	VerticalRangeType-fixm/3.0	
LocationPointType	BearingType	VerticalRangeType-fixm/4.3	
ExtendedMultiTimeType	FlightLevelType	AircraftType-fixm/3.0	
MultiTimeType	FrequencyType	AircraftType-fixm/4.3	
TargetMultiTimeType	GroundSpeedType	AircraftTypeType-fixm/3.0	
TimeSequenceType	HeightType	AircraftTypeType-fixm/4.3	
AirspeedChoiceType	IndicatedAirspeedType	CommunicationCapabilitiesType-fixm/3.0	
AirspeedRangeType	MassType	CommunicationCapabilitiesType-fixm/4.3	
AltitudeChoiceType	TrueAirspeedType	NavigationCapabilitiesType-fixm/3.0	
BeaconCodeType	WindSpeedType	NavigationCapabilitiesType-fixm/4.3	
GloballyUniqueFlightIdentifierType	FlightLevelOrAltitudeChoiceType	SurveillanceCapabilitiesType-fixm/3.0	
GroundspeedChoiceType	FlightLevelOrAltitudeOrRangeChoiceType	SurveillanceCapabilitiesType-fixm/4.3	
GroundspeedRangeType	TimeChoiceType	SurvivalCapabilitiesType-fixm/3.0	
LateralOfftrackType	TimeRangeType	SurvivalCapabilitiesType-fixm/4.3	
OfftrackDistanceType	TrueAirspeedChoiceType	DangerousGoodsType-fixm/3.0	
RadioFrequencyType	TrueAirspeedRangeType	DangerousGoodsType-fixm/4.3	
SpeedType	RestrictionReferenceType	ShippingInformationType-fixm/3.0	
AircraftCapabilitiesType	UniversallyUniqueIdentifierType	ShippingInformationType-fixm/4.3	
DinghyType	AircraftTypeChoiceType	AllPackedInOneType-fixm/3.0	

Only 2 complexTypes remained completely unchanged between v4.0 and v4.3

Other Approaches

Walmsley, Priscilla. "Comparing and diffing XML schemas."
Presented at Balisage: The Markup Conference 2015,
Washington, DC, August 11 - 14, 2015. In *Proceedings of
Balisage: The Markup Conference 2015*. Balisage Series on
Markup Technologies, vol. 15
(2015). <https://doi.org/10.4242/BalisageVol15.Walmsley01>.

Why was this was not my development experience?

The approach

- ◆ Use XSLT. (What else?)
- ◆ Use a pipeline (in Ant) to break it down into several simpler steps.
- ◆ Build it "iteratively". Which means:
 1. Write some code.
 2. Run it.
 3. Bang head against wall.
 4. Reread note to self.
 5. Goto 1.
- ◆ Current status: basic working code, much untested code, long list of future enhancements.

Steps

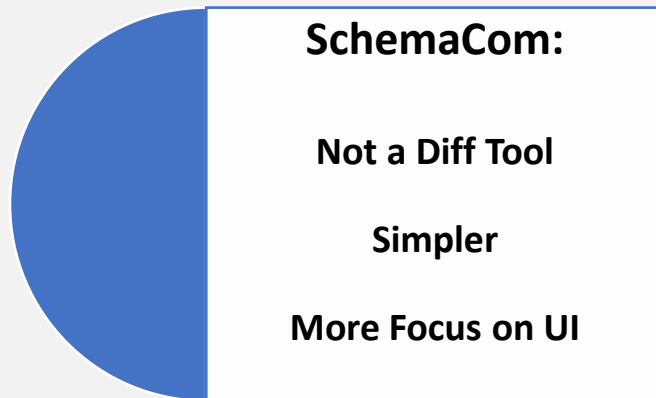
1. Canonicalize
2. Flatten
3. Simplify
4. Sort
5. Find Diffs
6. Determine Backward Compatibility
7. Report

SchemaCom only entails steps 2, 5 and 7

190 lines of XSLT split over 4 modules

120 lines of XQuery split over 2 modules

122 lines of CSS



Designed to assist in the derivation of mappings for vocabulary translation – only incidentally a diff tool.

Not concerned with subsetting or backward compatibility

Only tested on XSD 1.0

Compares versions which could consist of multiple schemas rather than physical schemas

Ignores factors that don't visibly impact the instance e.g changes to compositors, annotations and documentation.

Compares after inlining references to elements, groups and attributeGroups

Not concerned with changes to element sequencing as translation algorithm can be based on the sequence in the target vocabulary

Comparison is based on entity names. No attempt is made to compare anonymous types.

Design – Generate Schemas XML

AllSchemas.xsl (Standalone Step)

Place all the schema files into a parent folder for the whole vocabulary

<allSchemas>

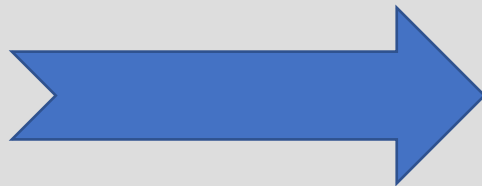
 <schemas version=...

 for each version in vocabulary

 for each schema in version

 inline groups and attributeGroups

 remove annotations documentation



```
<allSchemas>
  <schemas version="fixm/3.0">
    ...
  </schemas>
  <schemas version="fixm/4.0">
    ...
  </schemas>
  <schemas version="fixm/4.1">
    ...
  </schemas>
  <schemas version="fixm/4.2">
    ...
  </schemas>
  <schemas version="fixm/4.3">
    ...
  </schemas>
</allSchemas>
```

Design – Generate Comparisons

SchemaDeltas.xq

<deltas>

for each \$schemaPair in pairUpSchemas

fn:transform(complexDeltas.xsl, \$schemaPair)

```
<allSchemas>
  <schemas version="fixm/3.0">
    ...
  </schemas>
  <schemas version="fixm/4.0">
    ...
  </schemas>
  <schemas version="fixm/4.1">
    ...
  </schemas>
  <schemas version="fixm/4.2">
    ...
  </schemas>
  <schemas version="fixm/4.3">
    ...
  </schemas>
</allSchemas>
```



```
<deltas type="complex">
  <pairing versions="fixm/3.0 vs fixm/4.0" filename="pairings/complexfixm30vsfixm40.xml">
    <schemas version="fixm/3.0">
      ...
    </schemas>
    <schemas version="fixm/4.0">
      ...
    </schemas>
  </pairing>
  <pairing versions="fixm/4.0 vs fixm/4.1" filename="pairings/complexfixm40vsfixm41.xml">
    ...
  </pairing>
  <pairing versions="fixm/4.1 vs fixm/4.2" filename="pairings/complexfixm41vsfixm42.xml">
    ...
  </pairing>
  <pairing versions="fixm/4.2 vs fixm/4.3" filename="pairings/complexfixm42vsfixm43.xml">
    ...
  </pairing>
</deltas>
```

deltaForms.xq

**Generates
Xform UI**

Rendered
as xhtml

Delta files
from
previous
step

Bind to UI
controls

Xform
model



Reiterating
what made
this an
approachable
problem

Can choose to
ignore changes
that won't be
visible in the
XML instances.

Whatever has
been achieved
here is the result
of naively trying
to solve a similar
but different
problem.

We are aiming
for a best
guess and not
unerring
accuracy.

Future Work?

Performance degradation of UI when the size of the XForm model is large (say 1MB)

Adapting interface to allow the specification of mappings, however challenge to find space for additional UI interactions

Encoding a schema as a function of it's similarities and differences to another schema



Acknowledgements

Hans-Jurgen Rennau reviewed the paper, application and provided valuable and extensive feedback.