INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 8
MPEG GENOMIC CODING

ISO/IEC JTC 1/SC 29/WG 8 M60172
Online – July 2022

Title: **Linkage Metadata for Enabling Clinical Data Interchange**

Author: **Nidhi Tiwari (Royal Philips), Silvia Llorente (DMAG/IMP-UPC), Jaime Delgado (DMAG/IMP-UPC), Patrick Y.H. Cheung (Royal Philips)**

1. Introduction

The goal of this proposal is to provide a standard clinical data linkage Info that will make it easier to access data from multiple external data repositories associated with a specific sample, such as electronic health record (EHR) systems, hospital information systems (HIS) and laboratory information systems (LIS) which may be supporting different data exchange standards or vendor specific solutions, such as HL7, FHIR, Phenopackets, etc.

This solution will prevent data loss that can occur when a user is trying to manually retrieve data for a sample from external sources without following any standard thus maintaining data quality. User can access external data for a specific sample by referring to **ClinicalDataLinkageInfo** metadata. ClinicalDataLinkageInfo will contain information on data sources associated with a specific sample and specifications on how to access this data. The proposal also specifies data field mappings that is required when extracting data from external repositories following different protocols.

It will also help MPEG-G standard for genomic data storage and management. Adapting MPEG-G to accommodate well-established technologies in the healthcare IT ecosystem can broaden its appeal by reducing implementation costs and natively supporting a more diverse array of software systems. The seamless exchange of both clinical and genomic data is also a critical component for precision medicine informatics solutions. file formats, such as MPEG-G (ISO/IEC 23092 standard for the coding of genomic information), for retrieving additional clinical information for selected samples as needed.

Figure 1 shows an overview of our proposed generic clinical data interchange framework including different functional components and their relationships. It also illustrates how the framework can be integrated with MPEG-G servers, hospital systems and applications that require clinical data as inputs. Note that linkage information can be stored directly within an MPEG-G dataset as metadata or linkage attributes, rather than in a separate registry, to indicate the availability of additional clinical data in external repositories and the means to access them for specific samples. By making linkage information an intrinsic part of an MPEG-G dataset, it ensures the ready support for clinical data exploration and the linkage information will not be lost during file transport or system migration.

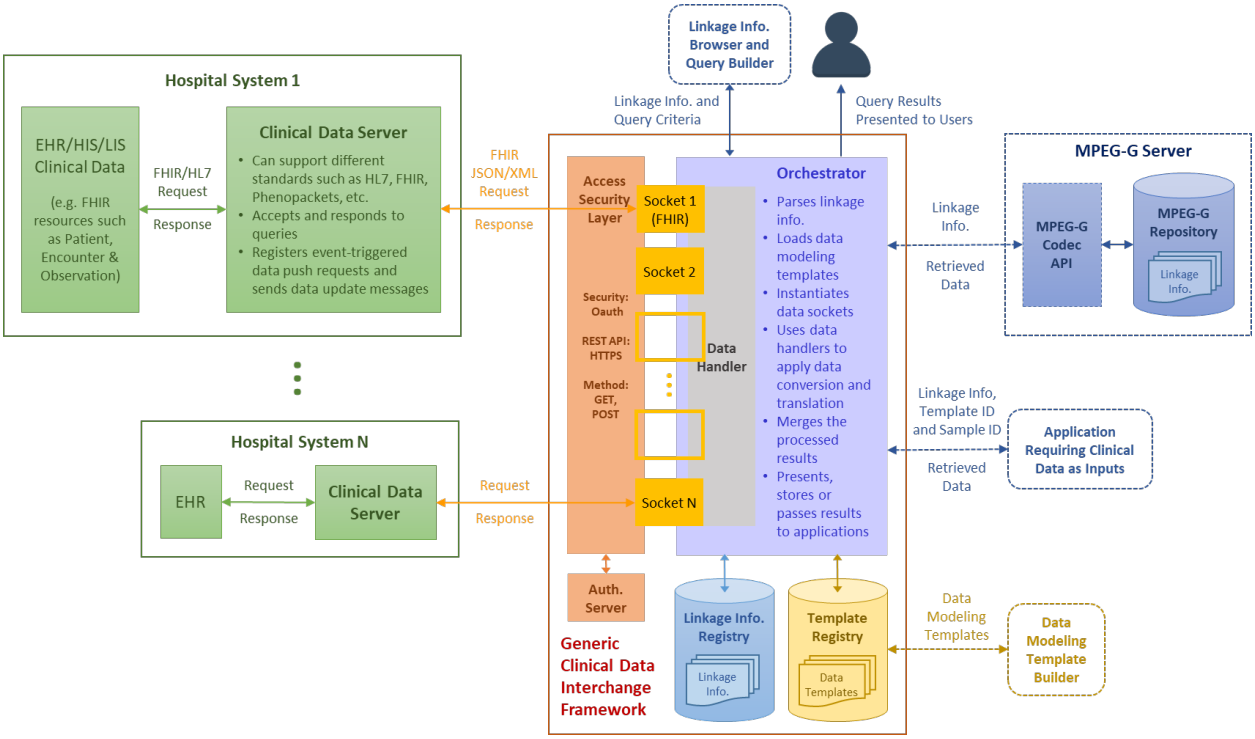


Figure 1 Overview of Proposed Generic Clinical Data Interchange Framework

Note that in Figure 1, a hospital system is depicted as consisting of two components: a backend EHR/HIS/LIS system communicating in some proprietary protocols, and a clinical data server that interacts with the backend system and connects it to the external environment. A clinical data server can be omitted if the EHR/HIS/LIS system can directly communicate with external applications using commonly accepted data exchange protocols. It provides the following functionalities:

- translating and reformatting the query/update/response messages between the internal and external protocols
- performing message validation and discarding any invalid messages
- keeping track of any active dynamic update requests of external clients, and in response posting push-in updates (e.g. a FHIR bundle) to the corresponding data interchange gateways upon receiving relevant event-triggered messages from the backend system
- authenticating and authorizing data access requests as needed

The following is a sequence of events for the query of clinical data using the proposed framework:

- Available linkage information is extracted from a registry or dataset, e.g., in MPEG-G format.
- User browses the linkage information that includes the availability of clinical data in external repositories for specific samples and the applicable data modeling templates.
- User selects the samples for clinical data retrieval and the data modeling template to apply.
- Orchestrator parses the linkage information associated with the selected samples and instantiates one or multiple protocol-specific data sockets, with linkage and template information dispatched, for handling communications with external repositories.
- Data Sockets generate query messages based on the protocols of the target servers and using relevant linkage and template information.

- Data Sockets parse query response messages received from external repositories.
- Orchestrator uses data handlers to apply the required data conversion and semantics translations on the results from data sockets as specified in the template, and merge them together.
- The merged results are presented to the user and/or stored in a dataset.

2. Clinical Data Linkage Metadata

Clinical Data Linkage metadata contains information about the associated local data entities for a given Sample and how to link them to multiple Data Sources. It consists of two main data elements: `DataSources` and `Samples`, and two optional `DataFieldMappingFile` and `DataFieldMappings` elements.

2.1 Data Sources

`DataSources` is a collection of `DataSource` elements, each containing the following elements:

- `Id`: unique identifier of the data source
- `Name`: name of the data source
- `URL`: URL of the data source
- `Protocol`: data exchange protocol such as HL7 and FHIR
- `Method`: GET, POST, PUT
- `Authorization` specific information related to the data source
    - `Method`: Authorization Method like oAuth2
    - `URL`: Auth Server URL to get access token
    - `Access`
        - `Type`: Auth Type like Client Credential, Asymmetric (public key) for FHIR
        - `AuthHeader`: Key and Secret required to get access token for Auth Type Client Credential. The IsEncrypted attribute for AuthHeader is set to true indicating that the values are in ciphertext
        - `Scope`: Resource Permissions provided by the server like Patient.Read, Observation. *
- `QueryFields`: Available Search parameters implemented by the data source that can be used to query the datasource to extract clinical data.

  For Example, `Sample` metadata may contain MRN to extract data from a remote clinical data repository, but the Data source may or may not have implemented MRN as a Search Criteria. Thus, <`QueryFields`> contains information about all Search parameters available to query that external data repository.

  In this scenario when MRN search parameter is not implemented on external data repository we may utilize other local entities to query the remote repository, like SSN.

- If `Sample` or `DataSource` metadata consists of field values of type “local” it means their corresponding mapping for different standards like FHIR, HL7 can be found in the `DataFieldMapping` template

An example `DataSources` from `ClinicalDataLinkageInfo` is available in Table 1.

Table 1 Example `DataSources` *Specified in XML Format*

```
<DataSources>
    <DataSource Id="DS0001" name="Data Source 1">
      <System>
         <URL>http://datasource1.com/</URL>
         <Protocol>FHIR</Protocol>
         <Method>GET</Method>
         <Authorization method="oAuth2">
           <URL>https://api.preview.platform.datasource1.com/oauth2/v1/token</URL>
           <Access type="Client Credential">
              <EncAuthHeader type="clientId">
                 <xenc:EncryptedData>
                   <ds:KeyInfo><ds:KeyName>TestKEYAuth</ds:KeyName></ds:KeyInfo>
                   <xenc:CipherData>
                      <xenc:CipherValue>YjIxZDMzYjAtYW05OS0y=</xenc:CipherValue>
                   </xenc:CipherData>
                 </xenc:EncryptedData>
              </EncAuthHeader>
              <EncAuthHeader type="clientSecret">
                 <xenc:EncryptedData>
                   <ds:KeyInfo><ds:KeyName>TestKEYSecret</ds:KeyName></ds:KeyInfo>
                   <xenc:CipherData>
                      <xenc:CipherValue>YmJDRmtEYXNNMkRLTC1KaYVU=</xenc:CipherValue>
                   </xenc:CipherData>
                 </xenc:EncryptedData>
              </EncAuthHeader>
             <Scope>datasource1/service/Patient.read datasource1/service/Observation.read</Scope>
           </Access>
         </Authorization>
      </System>
      <QueryFields>
         <Field Id="Patient.Identifier.type.SSN" type="external" />
         <Field Id="Patient.ID" type="external" />
         <Field Id="Patient.contact.telecom" type="external" />
         <Field Id="Patient.Identifier.type.DL" type="external" />
      </QueryFields>
    </DataSource>
    <DataSource Id="DS0002" name="Data Source 2">
      <System>
         <URL>https://datasource2.com/</URL>
         <Protocol>HL7</Protocol>
         <Method>TCP/IP</Method>
      </System>
    </DataSource>
</DataSources>
```

2.2 Samples

`Samples` is a collection of Sample elements, each containing the following elements:

- `Id`: Sample ID
- `AvailableDataSources`: a collection of `DataSource` elements, each specifying the Id of a data source containing clinical data for the sample
    - `Metadata`: consists of external field values associated with this `DataSource` that can be used to query the clinical data repository.
    - `Field` can be of two types external and local. `Field` has two attributes Id and type

- ➔ `Id` specifies the search parameter that can be used to query the datasource
- ➔ Field `type = "external"` indicates that the Field Id can be used directly to query the datasource
- ➔ Field `type = "local"` indicates that the Id attribute for this datasource is local and exact mapping information for specific protocol, associated with this datasource can be extracted from `<DataFieldMapping>` XML

- `Metadata`: a complex structure specifying the external data field values associated with this sample that can be used for query. It has an attribute IsEncrypted, if set to true, indicates that all the associated field values are in ciphertext. Otherwise, they are in plaintext. It is a collection of one or multiple field elements, each with an Id attribute specifying the external data field for query and the unique field value of this sample. It contains below attribute
- `IsEncrypted`: If it's true, then the `Metadata` value should be encrypted. If it's false, then the `Metadata` values are not encrypted.

  For regulatory compliance, IDs such as driver license number or SSN cannot be stored directly in the linkage information. To circumvent this issue, a hash value of the ID can be used as a universal key for the sample. Measures should also be taken to ensure that identifiable data of the sample cannot be returned by the clinical data repository.

  For storing user sensitive information `IsEncrypted` attribute for `Metadata` field is always set to true.

An example `Samples` from `ClinicalDataLinkageInfo` is available in Table 2.

Table 2 Example `Samples` *Specified in XML Format*

```
<Samples>
       <Sample Id="S0001">
            <AvailableDataSources>
                <DataSource Id="DS0001">
                    <Metadata>
                      <Field Id="Patient.ID" type="external">123445</Field>
                      <Field Id="Patient.contact" type="local">8578913099</Field>
                      <Field Id="Patient.Identifier.type.SSN" type="external">****29</Field>
                    </Metadata>
                </DataSource>
                <DataSource Id="DS0002">
                    <Metadata>
                      <Field Id="Patient.contact" type="local">8578913099</Field>
                    </Metadata>
                </DataSource>
            </AvailableDataSources>
            <Metadata>
                <EncField Id="MRN" type="local">
                    <xenc:EncryptedData>
                        <ds:KeyInfo><ds:KeyName>TestKEY</ds:KeyName></ds:KeyInfo>
                        <xenc:CipherData>
                            <xenc:CipherValue>aGVsbG8gd29ybGQ=</xenc:CipherValue>
                        </xenc:CipherData>
                    </xenc:EncryptedData>
                </EncField>
                <EncField Id="Observation.EncounterNumber" type="external" protocol="FHIR">
                    <xenc:EncryptedData>
                        <ds:KeyInfo><ds:KeyName>TestKEY2</ds:KeyName></ds:KeyInfo>
```

```
                        <xenc:CipherData>
                            <xenc:CipherValue>SSBhbSB0ZXN0aW5nIGNkaQ==</xenc:CipherValue>
                        </xenc:CipherData>
                    </xenc:EncryptedData>
                </EncField>
            </Metadata>
        </Sample>
        <Sample Id="S0002">
            <AvailableDataSources>
                <DataSource Id="DS0001">
                    <Metadata>
                      <Field Id="Patient.ID" type="external">123445</Field>
                      <Field Id="Patient.contact" type="local">8578913099</Field>
                    </Metadata>
                </DataSource>
                <DataSource Id="DS0002"></DataSource>
            </AvailableDataSources>
            <Metadata>
                <Field Id="MRN" type="local">MRN200</Field>
                <Field Id="PV1.EncounterNumber" type="external" protocol="HL7">EN200</Field>
            </Metadata>
         </Sample>
</Samples>
```

2.3 Data Field Mappings

`DataFieldMapping` contains mapping information for fields with type=" local" used in `Sample` or `DataSource` metadata against the search identifier used in an external clinical repository for a given protocol like HL7 or FHIR.

A local entity can have different mapping for different data exchange standards and all these mappings are included in the DataFieldMapping.

For example, if a `Sample` metadata field is MRN with type=" local" and that `Sample` extracts data from a repository that follows FHIR standard then the linkage dictionary i.e., `<DataFieldMapping>` template will contain mapping as below:

```
<FieldMapping localId="MRN" name="Medical Record Number">
      <ExtId protocol="FHIR">Patient.identifier.type.MR</ExtId>
</FieldMapping>
```

An example `DataFieldMappings` is available in Table 3.

Table 3 Example `DataFieldMappings` *in XML Format*

```
<?xml version="1.0" encoding="UTF-8"?>
<DataFieldMappings xmlns="urn:mpeg:mpeg-g:linkage:dataset:2022"
    xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:xenc="http://www.w3.org/2001/04/xmlenc#"
    xmlns:ds="http://www.w3.org/2000/09/xmldsig#">
      <FieldMapping localId="contact" name="Patient Contact">
         <ExtId protocol="FHIR">Patient.contact.telecom</ExtId>
         <ExtId protocol="FHIR">Patient.telecom</ExtId>
         <ExtId protocol="HL7">PID.contact</ExtId>
      </FieldMapping>
      <FieldMapping localId="SSN" name="Social Security Number">
         <ExtId protocol="FHIR">Patient.identifier.type.SSN</ExtId>
         <ExtId protocol="HL7">PID.SSN</ExtId>
```

```
      </FieldMapping>
      <FieldMapping localId="MRN" name="Medical Record Number">
         <ExtId protocol="FHIR">Patient.identifier.type.MR</ExtId>
         <ExtId protocol="HL7">PID.MRN</ExtId>
      </FieldMapping>
      <FieldMapping localId="EncounterNumber" name="Encounter Number">
         <ExtId protocol="FHIR">Patient.Encounter</ExtId>
         <ExtId protocol="HL7">PV1.Encounter</ExtId>
      </FieldMapping>
</DataFieldMappings>
```

3. Use Cases

3.1 Explore External Clinical Data from Multiple Data Repositories for a Given Sample

A user wants to extract information for a given Sample A from two external repositories, Cerner (https://fhir.cerner.com/authorization/authorization-specification) and AthenaHealth (https://docs.athenahealth.com/api) using linkage metadata.

For the above use case Clinical Data Linkage Info will include below information:

- Sample metadata like Social Security Number
- The `<Datasources>` associated with the sample, for Example, this sample will contain two DataSources, Cerner and Athena Health along with their Authentication Information and available query fields for accessing this data source.
- The query mapping for Cerner using FHIR protocol is `Patient.Identifier.type.SSN` and for AthenaHealth it is `Patient.SSN`. `<DataFieldMapping>` will contain all protocol-specific mappings associated with the data field.
- Using these three components: **`Sample`, `DataSource` and `DataFieldMapping`**, a user is able to access the requested data from both the repositories that may be using different data exchange standard.

3.2 Explore External Clinical Data for Patients on Clinical Trial

A user explores a dataset of gene expressions in an MPEG-G file that contains linkage information for specific samples with additional clinical data that could be retrieved from external repositories.

The samples of the dataset are patients under a clinical trial for determining if there exists a group of genes that can predict the treatment response three months after the administration of the first dose of drug based on their expression levels measured before the treatment.

Three months have passed after the gene expression data was generated and now the user wants to collect the latest clinical observations, including heart rate and blood pressure.

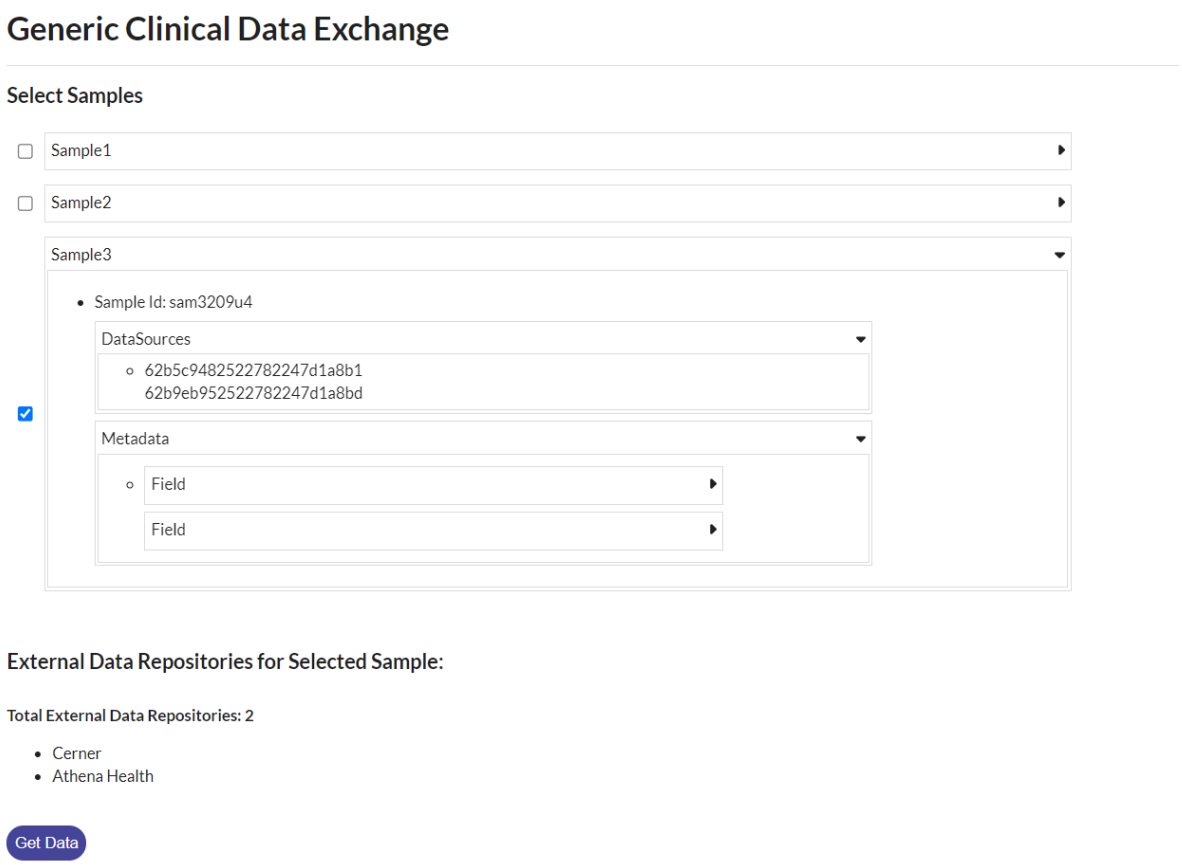
The user found from the Clinical Data Linkage Information Browser that all samples have clinical data available in external repositories

The user then selects all samples and submits request to multiple external repositories. Through Clinical Data Linkage Info, Data Source Info and Data Field Mapping a generic framework will be able to automatically access, parse, translate, merge, and represent the response as per user requirements.

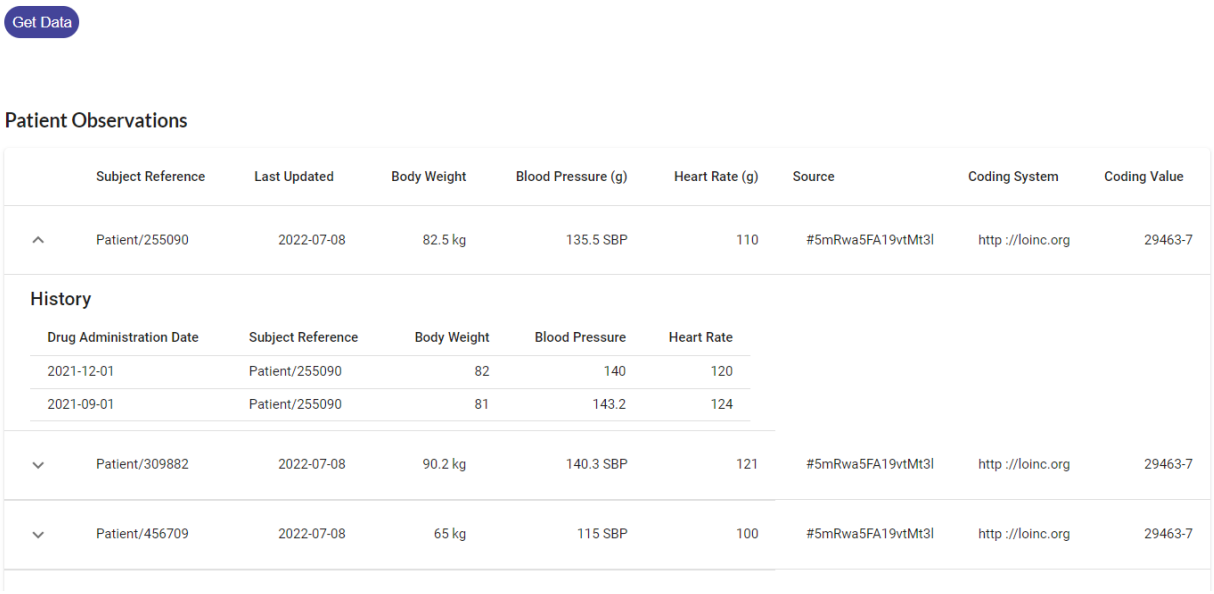
This results in the generation of the required clinical data of all samples, which are then stored as sample

attributes in the dataset using the MPEG-G codec. With the latest and accurate clinical data readily available to the user, it removes the burden for finding, downloading, and preparing the required clinical data, thus allowing the user to focus on real analysis.

Below snippet shows when a user selects a sample, it shows all external repositories associated with the sample and displays their names.



When User clicks on the button Get Data, the software with the help of **Clinical Data Linkage** metadata handles parsing, Authentication for different data sources supporting different protocols, conversion, translation and then displays Observation from both the external clinical repositories.



4. Conclusion

In conclusion, this proposal will standardize the clinical data retrieval from multiple data repositories that follow different standards for data exchange.

The linkage information can be stored directly within an MPEG-G dataset as metadata or linkage attributes, to indicate the availability of additional clinical data in external repositories along with specifications on how to access this data for specific samples. By making linkage information an intrinsic part of an MPEG-G dataset, it ensures the ready support for clinical data exploration and the linkage information will not be lost during file transport or system migration.

To test the functionality of the proposal in real time, a software prototype has been tested that utilizes **Clinical Data Linkage Metadata** to get data for a selected sample from two external data sources following different standards using `<ClinicalDataLinkageInfo>` and `<DataFieldMapping>` templates. The response from the clinical data repositories is readily available to the user to perform analysis and draw insights from the data.

Next steps would involve standardizing the Clinical Data Linkage Metadata and Data Field Mapping templates for multiple clinical data repositories and healthcare protocols. It includes more detailed research into building standard templates for DataSources and DataFieldMappings associated with the Samples.

Further a generic framework can be developed that utilizes the proposed solution and connects to multiple healthcare sources to access latest and precise data for analysis.

5. Supplementary Files

The following XML schema are included as supplementary files with this submission.

- instance-linkage.xml
- linkage.xsd
- instance-datafield-mapping.xml
- datafield-mapping.xsd