**Objective**

The main objective of LegalCiteM (LCM) is to provide a standard conceptual model to describe legal references that allows their representation into a number of different data formats.

**Terms defined**

A **textual citation** **(also “citation” or “source”)** is the way a source document describes a legal resource of some kind. This is usually done by mentioning relevant facts and data about the resource to help in identifying it univocally.

**A Reference** is a computer-usable representation of the same data specified in the citation.

A **Feature** in a reference is t named list of data values extracted from the reference. Features are meant to help identify and/or describe the resource referenced. All features are named to describe, qualify, and constrain their values. Features are always a \*list\* of values.

**Name** is a label that characterizes and contextualize the fact or data itself. For instance the value "2015" needs to be characterized by a name in order to be understood: is it a document number? A section? A date? If a date, an origin date or a version date?

**Values** are composed of zero, one, or many values (depending on the quantity of information provided in the citation). To simplify handling of data, therefore, all reference features are associated to a list that is ordered and arbitrarily long.

**Values: hierarchical or equivalent**

Values in a reference feature need to be identifiable in terms of **hierarchical** positions, wheresome values are more general/important/contextualizing than others. Or in terms of **equivalence**, where values are equivalent to each other.

Jurisdiction is often expressed as a list of hierarchical values. For instance, there are 41 places in the United States called Springfield, and therefore regulations for the city of Springfield in Alabama is differentiated from regulation for the city of Springfield in Colorado by the full specification of the hierarchy of jurisdiction, such as (United States, Alabama, Springfield) rather than (United States, Colorado, Springfield). On the other hand, official codes are often meant as equivalent to full-blown values, so for instance (United States, Alabama, Springfield) can be considered equivalent to (US, AL, Springfield) according to ISO 3166.

Similarly, languages are often expressed by codes, but, unlike jurisdiction, they are equivalent values. ISO 639 provides two separate lists of codes associated to human languages, a two-letter (ISO 639-1) and three-letter (ISO 639-2) list of codes that are differently used in current reference standards. Thus, we consider the list (fr, fra, française, French) as composed of **equivalent values**, each of which provides the same information as the others.

All values in a hierarchical list are relevant, and the order in which they are placed in the list provides a disambiguation path for the precise identification of the information. In contrast, any one value in an equivalence list is sufficient for disambiguating the information.

Depending on the amount of information provided in the citation, all hierarchical lists can be composed of a variable number of values. For instance, Italian act numbers are reset at the beginning of each year, so in order to identify an act one needs only year and act number, but it is customary to provide a full date: as such, the value list of a date can be a single year, a full date (year, month, day) or even a complete date-time specification (year, month, day, hour, minutes, seconds, time zone) depending on the amount of information provided in the citation.

**The** **Three Axes: Jurisdiction, Document Type, FRBR Level**

Features are described along three independent axes: the document types they characterize, the FRBR level they belong to, and the countries and jurisdictions that make use of them.

**Jurisdiction**: each country and jurisdiction has different ways to describe documents relevant to itself, which correspond to different features in their identification. While a few families of approaches can be identified (e.g., common law vs. civil law), differences exist and give ground to different features, different constraints on common features, and different assumptions on the existence of features. For instance, specifying a value for the language feature in a mono-lingual country such as the US has a completely different meaning (= "use the default") than in an intrinsically multi-lingual country such as Switzerland (= "any official language is acceptable").

**Document Type**: the LCM TC has identified four fundamental categories of document types: court documents, executive branch documents, legislation and parliamentary documents, and secondary material. Each of the four categories can and should be further refined in a complete catalog of document types. Each document type makes some features fundamental for identification, and others non-applicable or redundant. For instance, the emanating actor is often fundamental for executive documents, while it is often obvious for legislative documents.

**FRBR level**: each feature belongs to one and only one FRBR level. For instance, it is commonly agreed upon that the number is a work-specific feature, language is an expression-specific feature, and data format is a manifestation-specific feature. Each approved feature must therefore specify the FRBR level it applies to.

**The Source and Interpretation Frames**

LCM distinguishes the provenance of each feature by identifying two frames of reference: the **source frame** and the **interpretation frame**.

The **source frame** represents in a machine-readable way the same information that is specified in the source, e.g. the citation. The source frame represents exactly the information that the author of the source provided, regardless of how complete, correct or effective in identifying the resource it is. Obviously, sources may be incomplete, incorrect, ambiguous, and unable to identify precisely the resource.

The **interpretation frame** represents in a machine-readable way *additional* or *alternative information* that is the result of an active interpretation of the citation by the author of the reference, meant for disambiguating, correcting, or adding detail to the identification.

For instance given the citation "124 Stat. 119", the source frame will contain information only about the document type (Stat.), the volume number (119) and the page (124), while other information, such as the country (USA), the language (English) or the title ("Patient Protection and Affordable Care Act"), if appropriate or necessary for the identification, will need to be part of the interpretation frame.

**Representations of the Features**

LCM uses only two types of data structures for its organization: the **unordered list of name/value pairs called the "record"**, and the **ordered list of values called the "array"**. These data structures nest in a controlled way as explained in the following.

**Atomic values** can be strings, numbers, and booleans.

In all our examples we will borrow the syntax from JSON, using { } to represent records (objects in JSON parlance) and [ ] to represent arrays (called arrays in JSON, too). Names and strings are quoted—“”, while numbers and booleans are not.

LCM describes references using a data structure called a **feature list**. A feature list is composed of two **frames**--s**ource** and (optionally) **interpretation**, which are **records**. Each frame contains up to four FRBR levels, which are also records. Each level contains one or many features, which are arrays of hierarchical values. Each hierarchical value is an array of equivalent values, that are always atomic.

For instance the following is a representation of "124 Stat. 119" with some source values and some interpreted values. It is also, by construction, a valid JSON structure:

// frames

// FRBR levels

// feature

// hierarchy of equivalent values

{

 "source": {

 "work": {

 "document type": [

 ["Statute", "Stat."]

 ],

 "document number": [

 [119],

 [124]

 ]

 }

 },

 "interpretation": {

 "work": {

 "jurisdiction": [

 ["us", "USA"]

 ],

 "document title": [

 ["Patient Protection and Affordable Care Act"]

 ]

 },

 "expression": {

 "language": [

 ["en", "eng", "English"]

 ]

 }

 }

}

This feature list uses both frames. The source frame only uses the "work" FRBR level, and uses two features, "document type" and "document number". The interpretation frame uses both the "work" FRBR level and the "expression" FRBR level, containing respectively the "jurisdiction" and "document title" features, and the "language" feature.

The "document type" feature has only one hierarchical value, composed of two equivalent values, "Statute" and "Stat.". This means that consumer applications of the feature list are allowed to use whichever of these two values they prefer. The feature "document number", on the other hand, has two hierarchical values, each of which is composed of only ONE equivalent value. This means that the second value (124) is hierarchically dependent on the first value (119) so as to signify that the page number (124) is relevant in the context of the volume number (119) of the Statute.

The "jurisdiction" feature has one hierarchical values composed of two equivalent values, "us" and "USA", the "document title" has one hierarchical values composed of one equivalent values, and the "language" feature is composed of one hierarchical values composed of three equivalent values

More complicated examples can be imagined. For instance, the jurisdiction for a regulation for the city of San Francisco in California (United States) can be represented as follows:

 "jurisdiction": [

 ["us", "USA"],

 ["us-ca", "California"],

 ["San Francisco"]

 ],

which corresponds to a hierarchy of three terms, United States, the State of California, and the city of San Francisco. United States and California are both represented by two equivalent values each, the first being the ISO 3166 code for them, and the second a readable representation of the value. Equivalent values are never required, but can be specified for any value of any feature, *provided that the most important or codified atomic value is the first of the list*.

Dates are always stored as a hierarchy of elements in the following order: year, month, day, hours, minutes, seconds, time zone. Any reasonable combination of elements can be accepted, as long as it follows without gaps the order specified above. Thus acceptable lists are those composed of only the year, of year, month, day, of all elements but the time zone, etc.

**Linearizations of the Feature List**

LCM does not mandate any specific linearization of the feature list. Many existing linearization syntaxes, on the other hand, can be used to linearize feature lists with little or no changes to their use.

Some linearization syntaxes can only be used for specific classes of documents: for instance, ECLI can only be used for sentences of EU courts, while others are amenable to a wider range of document types. Some linearization syntaxes impose their own sets of features, and will require mapping towards the ones in LCM, while other do not predefine anything. For instance, OpenURL allow in principle any feature to be defined and used.

In principle, it is necessary that, for those documents that lie within the semantic reach of each linearization syntax, a clear and unambiguous mapping is provided so that the linearization syntax becomes a safe mechanism for representing LCM references. As such, it is a fundamental effort by the LegalCiteM TC to be able to identify exactly the semantic boundaries of each linearization syntax and the mapping from and to each others' feature sets. Given the wide overlap existing in the documents covered by each linearization syntaxes, this guarantees automatic translation, and therefore interoperability of document systems and document collections regardless of the syntax used for the references.

This mapping may make use of defaults and conversion rules in an automatic fashion. For instance, even if ELI does not mandate any specific syntax for the specification of the jurisdiction, this can be reliably deduced from the domain of the URI, and therefore the jurisdiction feature is automatically supplied even in the absence of a source feature with that semantics.

**Initial List of Features**

The full list of features to be dealt with depend on document type and local customs. Nonetheless, an initial and fundamental list of features can be provided, and is shown in the following. The names of the features as listed are not even a proposal, just a quick dump of the first term appearing in the conscious part of the brain.

schema ---- (e.g. eli, ecli, akoma ntoso, etc.)

jurisdiction work (a hierarchy whose first level is the country)

work author work (e.g. legislator)

work creation date work (a date, at least YYYY)

language expr (a code for the human language used)

format manif (a code for the data format used)

All allowed linearization syntaxes must either provide a feature that can be mapped to each of these LCM feature, or a default value that can be used instead.