

These are comments on

<http://www.unicode.org/reports/tr42/tr42-2.html>

1) The document defines a format for the representation of the UCD, but not the representation itself. Proposal: add to sec. 1

"Our representation supports the creation of documents that represent only parts of the UCD, either by not representing all the characters, or by not representing all the properties. This can be useful when only some of the data is needed."

the following sentence:

"This annex presents only the XML representation format of the UCD, but not the data in XML itself."

2) "Our schema defines a set of valid documents": change to "Our schema can be used for the creation and validation of documents". This seems to be a more appropriate description for what you are doing.

3) "It is important to stress that the Relax NG schema does not define valid documents.": change this to "It is important to stress that the schema which is defined in English propose imposes more constraints on the documents than what can be validated with the Relax NG schema.". That makes the relation between the prose text and the RELAX NG constraints (superset vs. subset) clearer.

4) "This is achieved by an inheritance mechanism, similar to property inheritance in CSS or in XSL:FO.": It would be good to have a definition of this mechanism and an overview of where it applies (e.g. in sec. 4.3 "Group"?), or at least an example close to this paragraph.

5) In Sec. 2.2 you say: "Non-Unicode properties can be represented by using elements in another (possibly empty) namespace, and/or by attributes in a non-empty namespace. Such elements are ignored for the purpose of determining the validity of a document.". Does this constraint apply to your schema, or is it a prose constraint? Note that your current RELAX NG schema forbids the use of elements and attributes from other namespace, so it contradicts the prose description. You could implement this constraint in the RELAX NG schema by adding the following definitions to it:

```
namespace ucd = "http://www.unicode.org/ns/2003/ucd/1.0"
```

```
namespace local = ""
```

```
anyElementExceptUcd =
```

```
element * - ucd:* {
```

```
  (attribute * { text }
```

```
  | text
```

```
  | anyElementExceptUcd)*
```

```
} *
```

```
anyAttributeExceptUcd =
```

```
attribute * - local:* { text } *
```

and add to each element declaration "anyElementExceptUcd" and "anyAttributeExceptUcd".

6) "# default; datatypes xsd = "http://www.w3.org/2001/XMLSchema-datatypes" is not necessary (though it also does not hurt).

7) Sec. 2.4 says "The root element of valid documents is a ucd.", but the example in sec. 10 does not use <ucd>.

8) I think you should decide between the two definitions for code-point given in [code point, 9] and [code point, 11]. Otherwise people will be confused while reading ucd documents and will have a harder time in writing them, or writing e.g. XPath expressions for their processing.

9) "element block { attribute first-cp { text }, attribute last-cp { text }, attribute name { text } } *" should be

"element block { attribute first-cp { text }, attribute last-cp { text }, attribute name { text } }+", so that there is at least one <block> element in a <blocks> element (this is what the prose says).

The same comment holds for the declarations of "element named-sequence" in "element named-sequences", and the elements in "element provisional-named-sequences", "element normalization-corrections", "element standardized-variants".

10) for the readability of the document, it would be very helpful to have links from referenced patterns to their definitions, e.g. a link to "set-of-code-points" and "code-point.content" from code-point |= element code-point { set-of-code-points, code-point.content }

Best regards,

Felix