An XML schema for taxonomic literature – taXMLit

Anna L. Weitzman[1] *& Christopher H. C. Lyal* [2]

[1] *The National Museum of Natural History, Smithsonian Institution, Washington D.C., 20560, U.S.A.;* [2]*The Natural History Museum, Cromwell Road, London SW7 5BD, UK*

The schema outlined below – 'taXMLit' - is designed to accommodate taxonomic literature. It covers all of the components of taxonomic publications and the taxon treatments contained within them other than the actual characters, which are dealt with by other projects. This explanation is written in conjunction with version 1.3 of taXMLit, which should be consulted for additional information. Many of the elements in the schema are annotated, and these annotations are not always reproduced below, although they are present in some of the figures.

Both taxonomists and those who need taxonomic information require greater access to material held in natural history museums and similar large biological repositories and their libraries. These repositories hold a wealth of inadequately accessible resources that describe and explain the diversity and depth of life on earth. Mining these data for research, conservation, drug discovery, protected area management, disease control, education, enjoyment of the natural world, etc., is difficult, time consuming, and often leads to redundant efforts. What should be a seamless, open "book" of knowledge consists, instead, of disparate, unintegrated sets of data - some in electronic form but most still on paper, and both published and unpublished.

Information held in museums centers on the following types of biological datasets: specimen collections, taxonomic databases, published taxonomic literature, geographical information systems, and unpublished archival materials. Making these information sources available is part of a larger, worldwide effort to enable easy access to the complete range of data required to understand individual species and their environmental and evolutionary relationships. This will require the establishment of cross-linkages between, and simultaneous access to, datasets from such information sources throughout the world. TaXMLit is one step needed to implement that vision.

The first use of the taXMLit schema will be to capture the text of the *Biologia Centrali-Americana* (BCA) (http://www.sil.si.edu/digitalcollections/bca). However, it will subsequently be used for other taxonomic works. It has been written with a focus on both botanical and zoological taxonomic literature and should also accept fungal and paleontological publications, but this has yet to be tested. The schema does not take into account the kinds of data needed for viral or bacterial publications.

Implementation of the schema will allow the text of the BCA to be fully searchable, and facilitate, in the first instance, static links from the text to other data sources (e.g. specimen databases on the web). The use of the schema for additional taxonomic works will allow links between different works that include the same taxa or their synonyms. Moreover, this opens the way for virtual compilations of taxon treatments to be made up by the user, comprising components of more than one original work, e.g. checklists, faunas, and floras. Another function

will be dynamic links to other kinds of databases as described above. These functions will require that the schema should, in the appropriate parts, be using the same or similar elements to schemas used by other relevant partners, and certainly be mappable to them.

TaXMLit is designed to hold text from the literature as it is presented, and not include interpretations of that text. For example, although locality names may be outdated (e.g. 'Burma' instead of modern-day 'Myanmar'), or subsequent information indicates that the cited publication date is incorrect, such information is not accounted for in the current schema. To accommodate this and similar information an 'interpretation layer' (TIL) will be created in a subsequent phase of the project. The TIL will function as both a simple layer around the publication that will allow authors to contribute information such as that above, and as a complex proxy layer between the digitized publication and other digitized information sources, including a variety of kinds of authority files. This facility will enable data to be entered independently after the digitization, in the same manner that any piece of text or data requires and receives interpretation in the normal course of use. The TIL will also allow interpretations to be attributed and dated, and allow for multiple interpretations of the same information by the same or different people or publications. Another function of the TIL will be in facilitating linkages between different taxonomic treatments, and between the treatments and other data sources. While the taXMLit uses elements that cover the same concepts as those used in other schemas (e.g. ABCD[1], designed for specimen data), the individual elements are not exactly the same, for a number of reasons. The TIL will facilitate mapping and linkage between schemas.

The taXMLit schema currently omits a number of generic kinds of metadata elements, such as those related to record creation, rights, etc. Developers of other standards such as ABCD & SDD have put a great deal of work into this in and we assume that the standards can be taken from those.

We currently plan to store the text of the BCA in database form, although there is no reason why taxonomic literature should not be stored as an XML document (given constraints on search speed with large XML files). We have designed the schema to accommodate both database and XML storage.

As it stands, the schema does not accommodate text formatting and structure or, for example, include page numbers. There are standard XML DTDs that handle these very well, we plan to use TEI-LITE to accommodate these components, and once the taXMLit schema is ready for use, we will combine the two.

The schema has been compiled in XMLSPY©. XMLSPY has also been used for testing the schema by entering data from a variety of taxonomic literature sources. Entry of text in trials showed us that a text string cannot be placed into an element which also has an attribute. Consequently, all elements containing text strings are simple types with no attributes, child elements of a complex type with one or more attributes (Fig 1.). For simplicity, in most of the figures below the simple text-containing element is omitted; the 'terminal' elements shown being the complex parent. Some terminal elements figured are 'complex types' which need further explanation; this is indicated in the 'type' box on the element.

[1] Access to Biological Collection Data - a joint CODATA and TDWG initiative. (http://www.bgbm.org/TDWG/CODATA/)

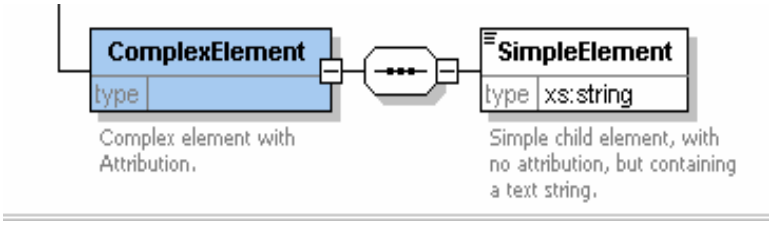


Fig. 1. General format of complex parent element and simple text-containing element.

Fig. 2. Types used in schema.

element	TaxonomicPublication
complexType	AbstractType
complexType	AcceptedOrValidTaxonNameType
complexType	AcknowledgementsType
complexType	AppendixType
complexType	BibliographyType
complexType	CitationType
complexType	ContentsType
complexType	ContributorType
complexType	DescriptionsType
complexType	DiscussionsType
complexType	DiscussionParagraphType
complexType	DistributionAndOrSpecimenType
complexType	ErrataType
complexType	FrontispieceType
complexType	GatheringType
complexType	GlossaryType
complexType	ImageCaptionType
complexType	ImageCrossReferenceType
complexType	IndexType
complexType	IntroductoryType
complexType	KeyDecisionNodeType
complexType	KeyToTaxaType
complexType	KeyWordsType
complexType	LocalityType
complexType	NomenclaturalTypeType
complexType	NomenclaturalTypeSpecimenType
complexType	NomenclaturalTypeTaxonType
complexType	NotesType
complexType	OtherCitationAuthorsType
complexType	PostscriptType
complexType	PublicationContributorType
complexType	PublicationDetailsType
complexType	PublicationSubHeadType
complexType	RelatedSpecimenType
complexType	RelatedTaxonInformationType
complexType	SpecimenType
complexType	SynonymCitationType
complexType	TaxonAuthorType
complexType	TaxonHeadingType
complexType	TaxonNameType
complexType	TaxonTreatmentType
complexType	VernacularNameType

Paragraphs of text in the original work are captured as a whole to facilitate the order of the text components to be captured and subsequently reconstructed. These paragraphs are usually also parsed into more or less detailed elements, as will be explained below. Each paragraph text-containing element (or rather, the complex element whose child contains the text string as it appears in the original) is given an ElementID. In addition to facilitating text reconstruction it also allows the use of an IDREF attribute elsewhere in the marked-up text, for example with dates, key lugs and images, to make the referred information available where required. The ElementID is applied to each paragraph (i.e. any text component terminated by the stroke of an 'Enter' key); in instances where there are multiple paragraphs in a section (such as descriptions) the containing element is unbounded and each paragraph has its own ElementID. ElementIDs run sequentially through the text.

Throughout much of taxonomic literature abbreviations are used (e.g. for genus names) or descriptors are omitted (e.g. for hierarchical rank levels above genus, or for components of label data if these are repeated). While this information is simple to derive for a human reader it is less accessible to machine treatment or amenable to database storage. For this reason many elements in the schema have the attribute 'Explicit', to denote whether the information included is explicitly stated or

implicit and derived either by programming code or by a human in the final verification of the markup. In all cases, these elements may be derived from the text content itself and are needed to build a complete database of the text that can be searched and made interoperable.

The schema comprises a number of complex Type elements (Fig 2), which will be explained below.

The root element of the schema is the TaxonomicPublication (Fig. 3). This can be any kind of publication, including multivolume works, articles in journals, and books. It has an attribute of 'TaxonomicPublicationID'. In order to accommodate publications that appeared in multiple volumes or fascicles there is a child element 'IndividualPublication'. We have separated the PublicationFrontMatter and PublicationBackMatter from the PublicationTaxonomicMatter, which contains the bulk of the taxonomic components. There is, of course, overlap between front and back matter, given that some publication components (e.g. glossary, acknowledgements) may occur either before the taxonomic content of a work or after it. Consequently, some elements appear (optionally) in both PublicationFrontMatter and PublicationBackMatter. The PublicationTaxonomicMatter is unbounded (may repeat as a group), since within some publications (such as the BCA) there are several components each with a separate author, title and even introductory matter ('PublicationTaxonomicSubhead'), which do not neatly lie within separate fascicles or volume parts. Using the method adopted allows us to take account of this.

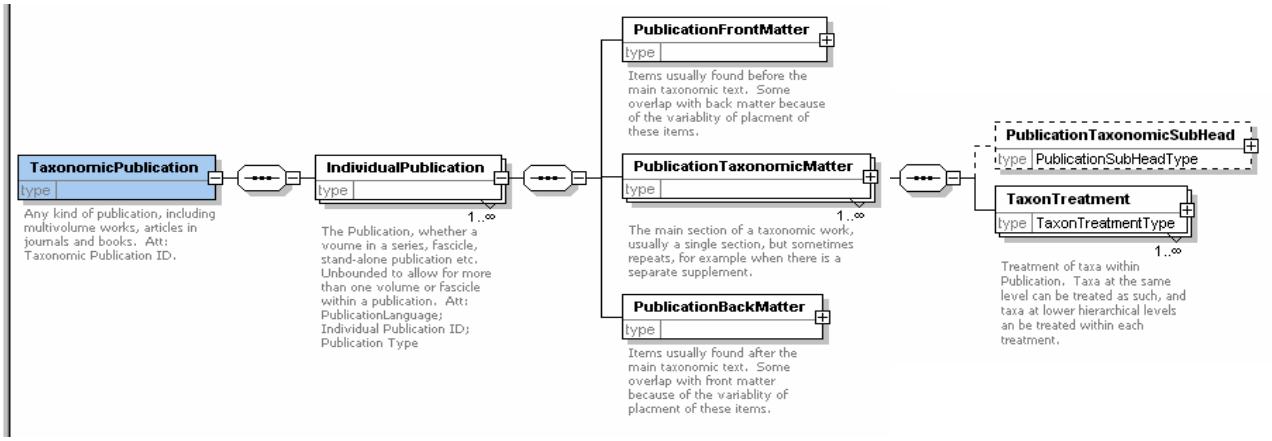


Fig. 3. Root element and basic publication contents.

Within the Types there are some repeating patterns. For example, the break-down of elements for names of persons (e.g. authors, editors, collectors) is always of the same pattern, as exemplified by the ContributorType (Fig. 4).

The full string of listed contributor(s) is stored in the ContributorString element. The atomized author names are stored in the 'ContributorAtomised' element, having the attribute 'OrderOfContributors'. A similar formulation is used for authors elsewhere in the schema, although with different attributes according to whether the author is of a taxon or of a publication. An ElementID is rarely required for contributors, since the contributors are usually part of another paragraph.

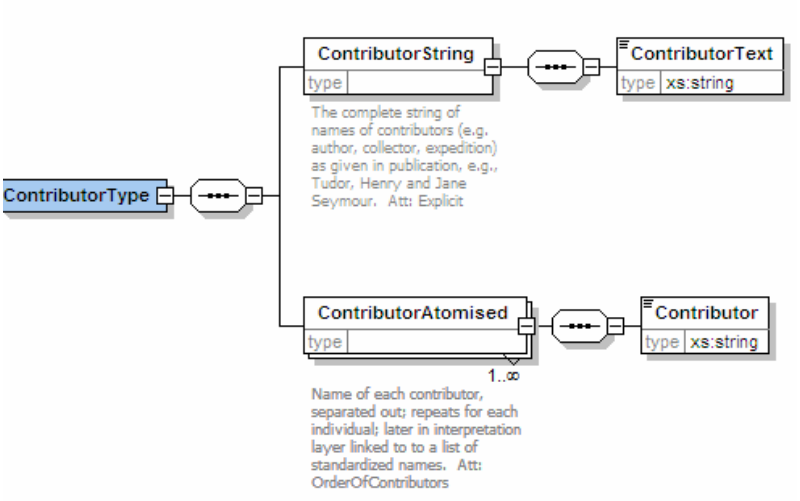


Fig. 4. ContributorType complex element

Within the PublicationFrontMatter and PublicationBackMatter elements are identified for each component of the text, some unbounded to accommodate them being used more than once. For example, 'IntroductorySection' (Fig. 5) is unbounded because it may be repeated for several different kinds of sections (e.g. Introduction, Foreword, Preface, Dedication) and the same format is suitable for each of these. 'Postscript' in the Back Matter is of the same format, and is similarly unbounded. 'Index' is also unbounded, since there may be more than one index within a single publication.

ElementIDs have been applied to each 'paragraph' of the IntroductoryMatter, as explained above, so that in the IntroductoryType the IntroductoryTitle and the IntroductoryAuthorPlaceDate text elements each have and ElementID. The IntroductoryParagraph (the text of the Introduction, Foreword etc) is unbounded, and each paragraph will have its own ElementID.

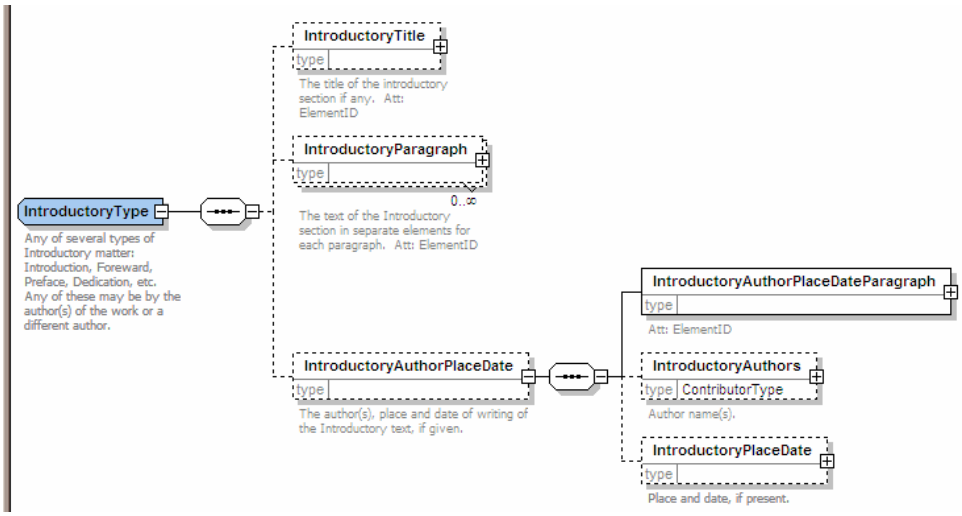


Fig. 5. IntroductoryType complex element. The elements holding the text strings have been omitted, as have the contents of the IntroductoryAuthors complex element, since this is the same architecture as Fig. 4.

Other elements of the PublicationFrontMatter and PublicationBackMatter are simpler than the IntroductoryType, in that their authors are the same as the rest of the publication. For example, the ContentsType (Fig. 6) comprises a heading, introductory text and the body of the contents itself. In this case, however, the Contents are often laid out in tabular form, so the format of the

ContentBody may be as text or table. Similarly the glossary, index and errata may be either format. In these cases the final schema incorporating the TEI-LITE elements will be used to provide the appropriate format.

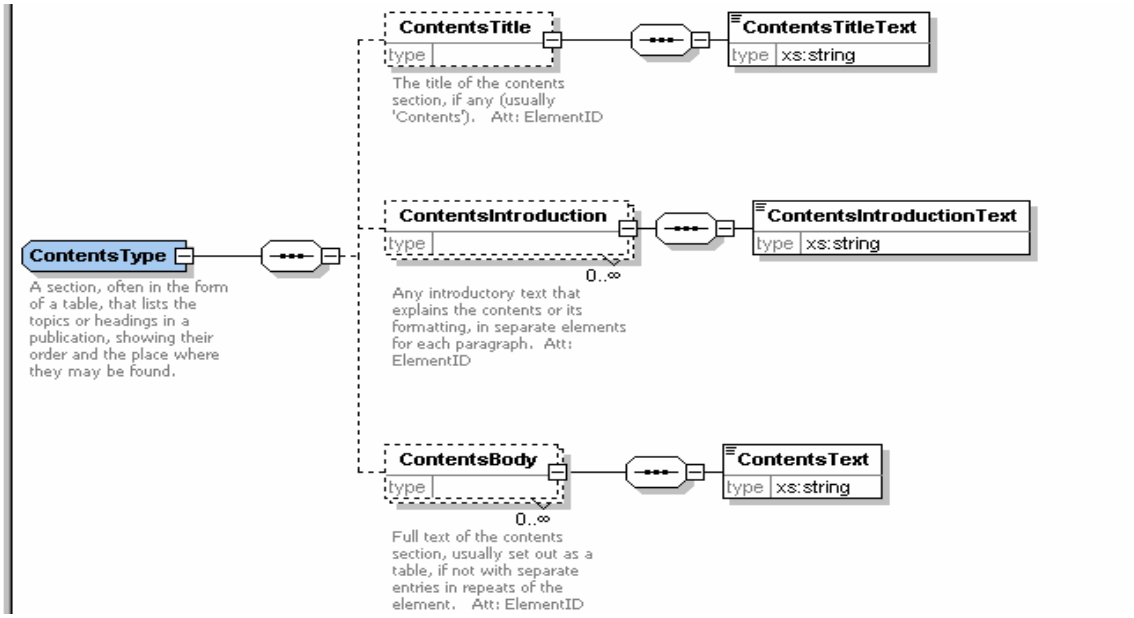


Fig. 6. ContentsType

Of the PublicationBackMatter, perhaps the most complex element is the BibliographyType (Fig. 7). The Title and Introduction are the same as those types described previously (and each possess ElementIDs), but the citations themselves are complex and often contain more detail than other citations used in taxonomic literature (which are enumerated in the PublicationDetailsType, Fig. 17). The BibliographType also includes yet a further type, the ImageCrossReferenceType (Fig. 34).

The main recursive unit of the schema is the TaxonTreatmentType (Fig. 8). As with most of the schema, almost every element is optional, since taxonomic literature, despite its strong structure, is very fluid as to what it actually contains. The TaxonTreatment element is used for each and every taxon included in a taxonomic publication. It has the attributes TaxonID; ParentNodeID; SiblingNodePreviousID; SiblingNodeNextID; RecognizedInTreatment; TreatmentLanguage. The four ID attributes will serve to place the treatment within the recovered text, and give the taxa a hierarchy. This is also possible through the recursive nature of the element, but in this case the function is to enable recovery from XML rather than database storage.

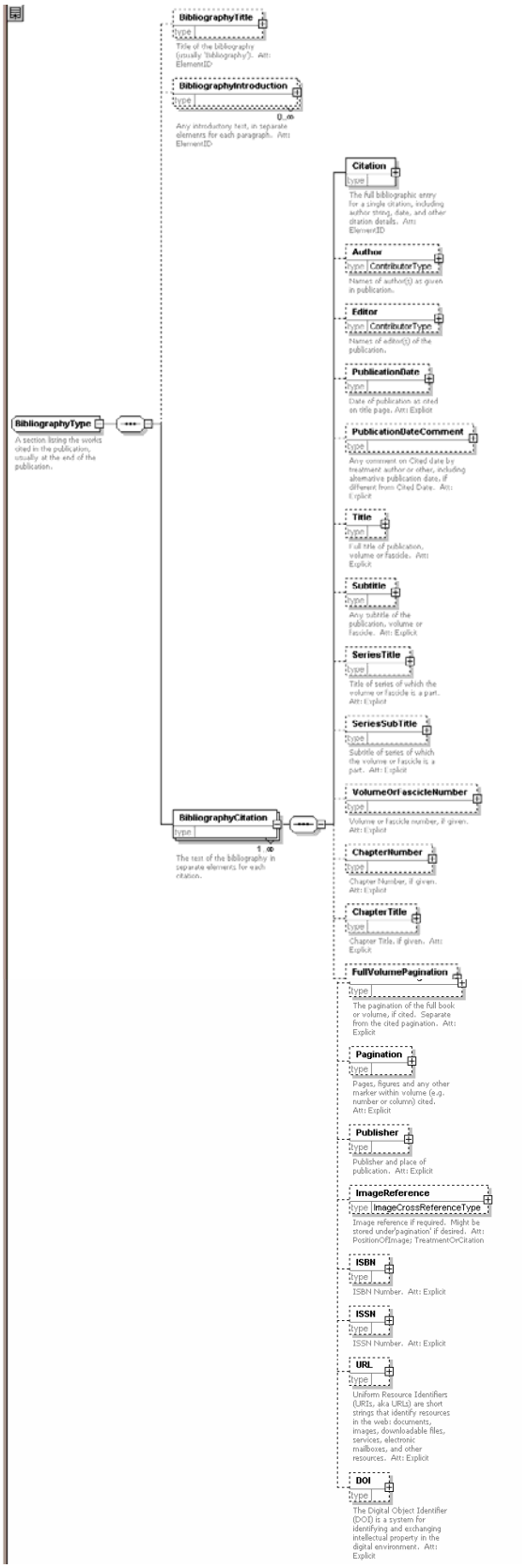


Fig. 7. BibliographyType complex element. The elements holding text strings have been omitted, as has the contents of 'Author', since this is the same architecture as Fig. 4.

The TaxonTreatmentType contains a TreatmentDate. This is the publication date of the taxon treatment according to information in the publication. This may differ from the date on the title page, as the latter reflects the entire publication, which may have appeared in parts over a span of time. The element is unbounded, allowing multiple dates to be entered. For example, some treatments are published in parts, e.g. in more than one fascicle, and in such instances the dates of each fascicle may be included. If there is no date present, then this information will have to be elsewhere, i.e. in the TIL. For the BCA the date is found as fascicle dates at the foot of some pages. The element has the attribute of IDREF (linking to the fascicle dates, if needed).

In many cases there is a hierarchy of names present placing the taxon that is the subject of the treatment. These are not taxon treatments themselves, but need to be included. This is done via an unbounded complex element the TaxonHierachyAbove (Fig. 9).

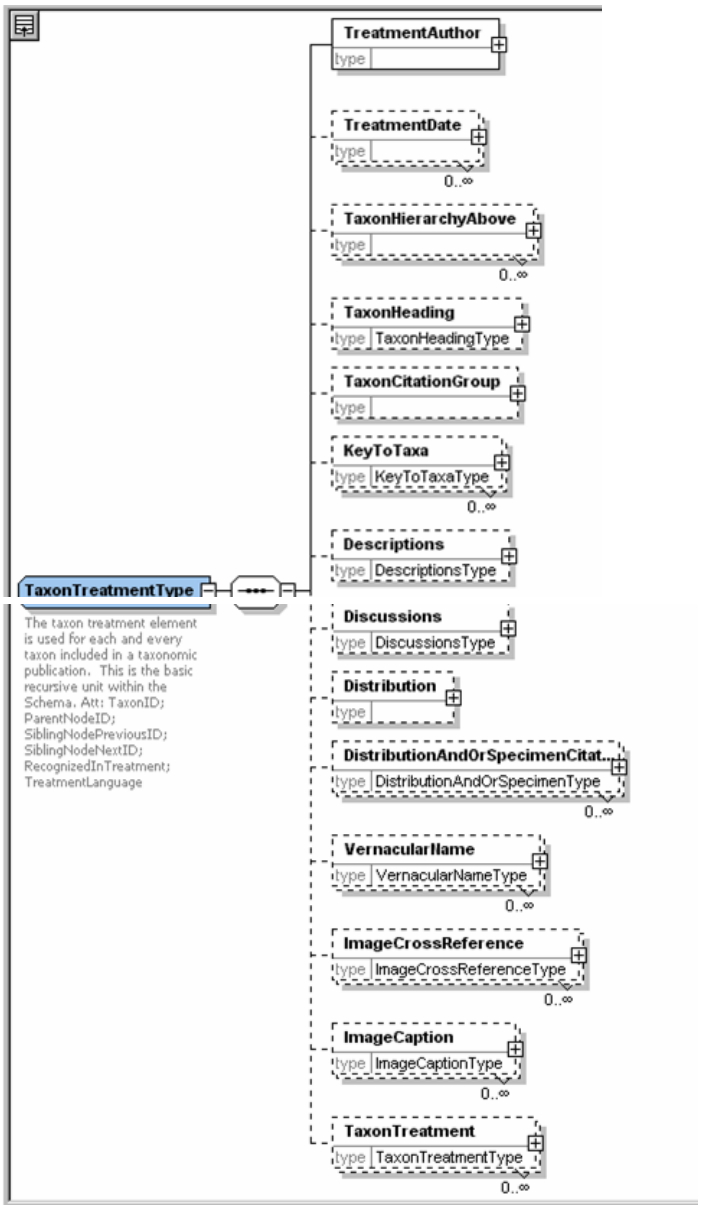
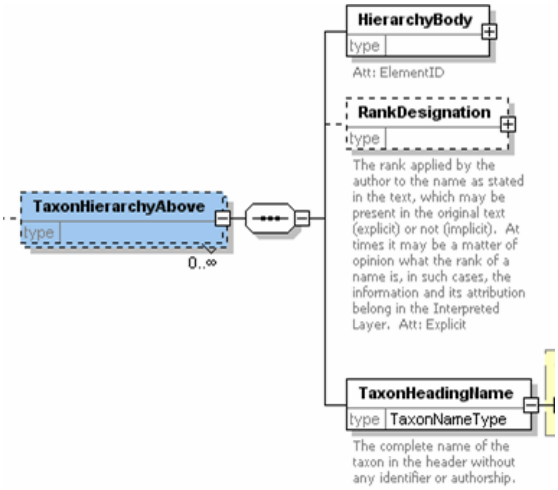


Fig. 8. TaxonTreatmentType

The TaxonHeading deals with the heading line of the taxon treatment, primarily the name and associated information (Fig. 10). As usual, most of the elements are optional. The first element, the TaxonHeadingBody, is the full text of the TaxonHeading, and carries the ElementID attribute. This attribute is then unnecessary elsewhere within TaxonHeadingType. The other elements contain the atomized data from the Heading text, including elements that might be stated explicitly or be implicit. An example of implicit data is in the taxon author, which may be stated (especially if different from the treatment author) or, if the same as the treatment author, simply be omitted, and therefore be implicit. Within the TaxonHeadingType, RankDesignation is the rank applied by the author to the name as stated in the text, which may be present in the original text (explicit) or not (implicit). At times it may be a matter of opinion what the rank of a name is and, in such cases, the

Fig. 9. TaxonHierachyAbove complex element. The elements holding text strings have been omitted.

information and its attribution belong in the TIL. Similarly, any reassessment or realignment of rank or relative rank order belongs in the TIL. The TaxonIdentifier is used for any identification numeral or letter that identifies the taxon within the heading text. Usually this takes the form of consecutive numbering for names with the same rank within the taxon of next higher rank.



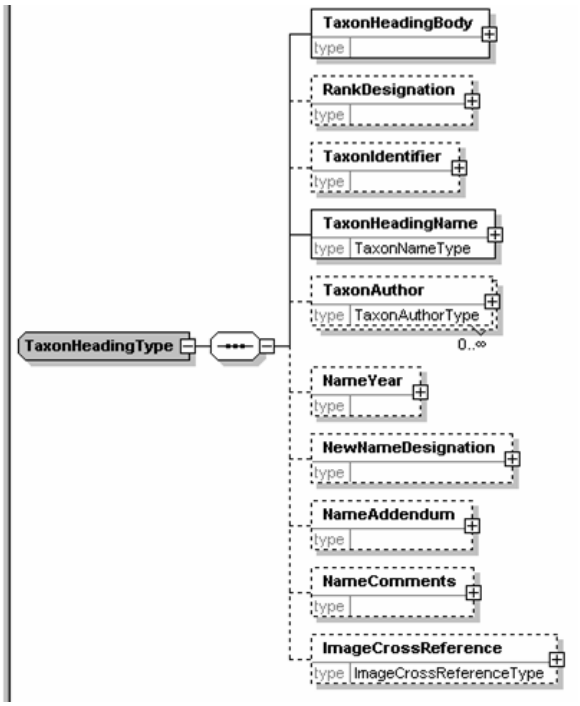


Fig. 10. TaxonHeadingType. The elements holding text strings have been omitted.

The TaxonAuthor element contains the stated Taxon Author(s) as distinct from the Treatment author(s). The attributes of this element are 'Explicit' and 'AppliesToRank', for those cases in which more than one author string is included in a name string that applies to the name at more than one rank (e.g. species and variety). The TaxonAuthorAtomised also has the attributes KindOfAuthor to distinguish between the kinds of roles that taxon authors may have (i.e. Basionym [Parenthetic], vs. Original [or Combination] Author(s), and, if applicable, 'in', 'manuscript' or 'apud') and Order of Authors. The NameYear is the year of publication as (if) given in heading. The NewNameDesignation is for text where the author believes he/she is erecting a new name, new combination, or giving a new status; e.g. "nom. nov.", "sp. nov.", "gen. nov.", "comb. nov.", "stat. nov." etc. The NameAddendum element is for any comment following the name other than the authors' comments regarding its status or certainty, whilst the NameComments element is for authors' comments regarding status or level of certainty attributed to name/rank combination, e.g. "cf." "aff.", "?".

The more complex Types that occur in the TaxonHeadingType are TaxonHeadingName (an instance of TaxonNameType) and ImageCrossReference (Fig. 34). The first of these, TaxonNameType is shown in Fig. 11.

The TaxonName element holds the full name string. It has an attribute of an InformalName flag, indicating whether the name is purported to be a valid (Bot) or available (Zoo) name.

The Genus name is required for any name of rank genus or below. As with SpeciesEpithet, InformalName, Breed and NamedIndividual, this is a simple element, containing only the text string.

The InfraGenericGroup (Fig. 12) is a little more complex than others in this Type. The name and rank are grouped to allow infrageneric names at different ranks. The element is repeatable to allow for cases where there are multiple occurrences of infrageneric names in the name string. For the same reason, it has the attribute 'OrderOfNames'. The InfraSpecificGroup has exactly the same structure and attribute, for the same reason.

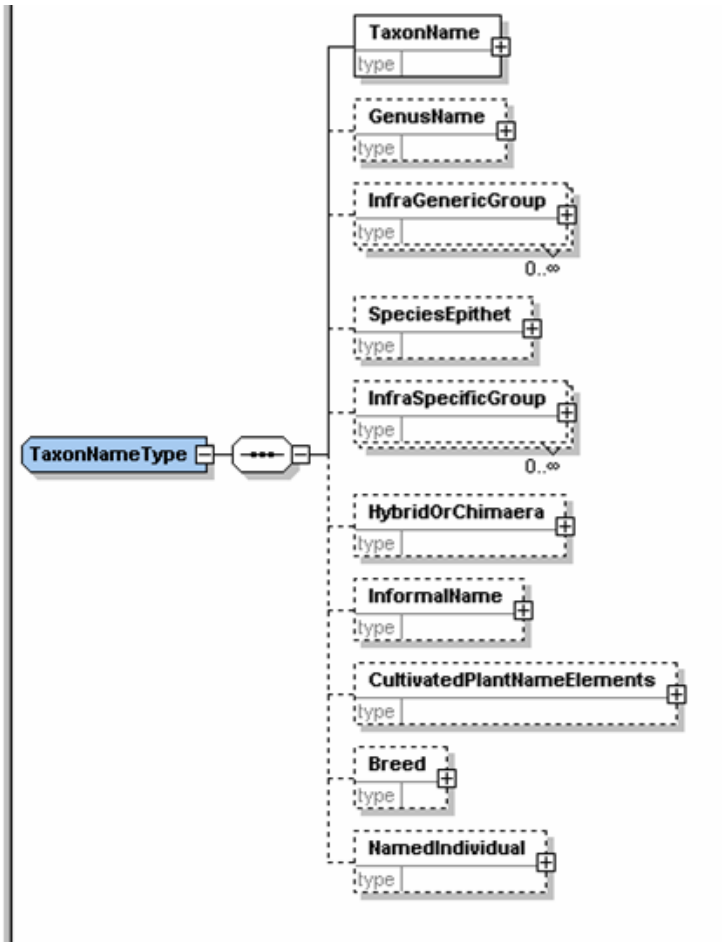


Fig. 11. TaxonNameType complex element. The elements holding text strings have been omitted.

The most complex of the elements within the TaxonNameType is the HybridOrChimaera element. This is presented in Fig. 13. It is a designation and atomisation of the elements needed for hybrid or graft-chimaera names or formulas. Named hybrids or graft-chimaera and hybrid formulas can both be accommodated. The container element itself has the attribute 'HybridStatus'. For named hybrids or chimaera there is a rank element to accommodate the rank at which the hybrid or graft-chimaera is noted, and a SymbolOrWord element for the symbol, letter or word used to designate the hybrid or graft-chimaera (i.e. 'x', '+', 'notho-' or 'n-'). The HybridFormulaAtomisation element allows for hybrid or graft-chimaera

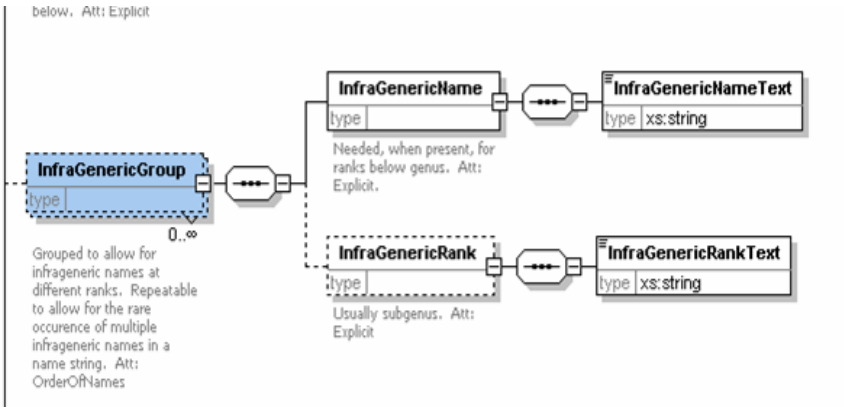


Fig. 12. InfraGenericGroup complex Element.

parents' names to be atomised. The first parent is taken to be the name in the TaxonName element, so this is not repeated here; the sex of each parent may be captured. The AdditionalParentNames container element includes a subset of the TaxonNameType, the corresponding elements having the same internal structure.

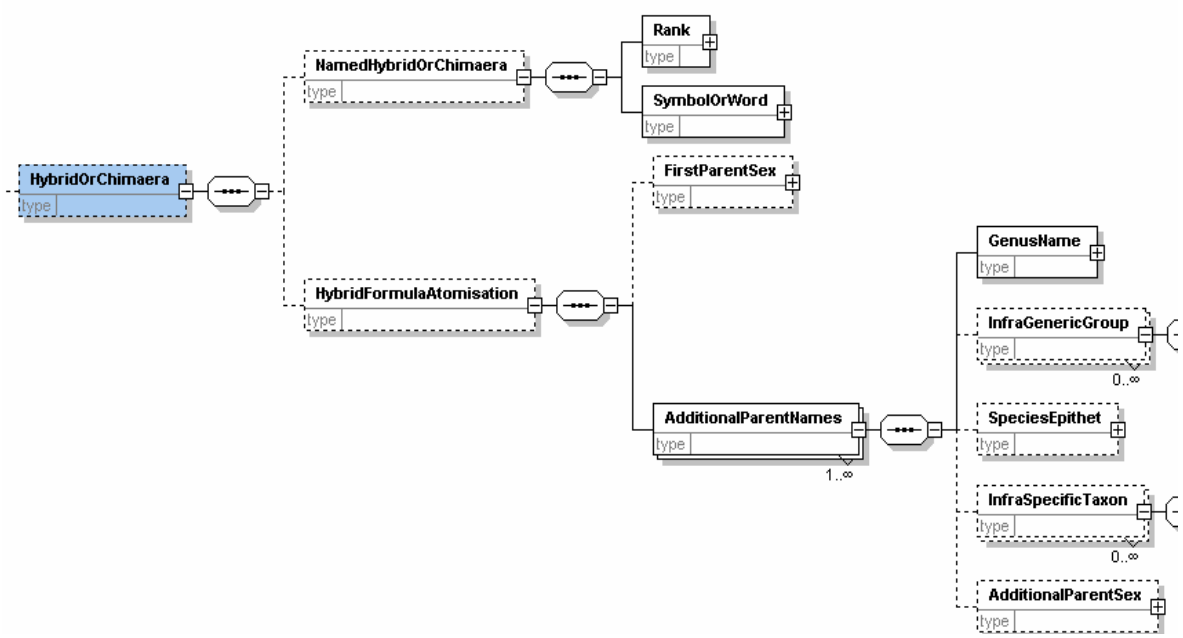


Fig. 13. HybridOrChimaera Complex Element. The elements holding text strings have been omitted.

The next major section is the TaxonCitationGroup, which contains the accepted/valid name of the taxon, with the appropriate references, and the synonyms and other names, again with citations. These two sections are treated as separate subsets, as shown in Fig. 14. The AcceptedOrValidNameType (Fig. 15) includes many of the elements of the TaxonHeadingType, along with other information likely to be cited, including publication, other citations etc. The first element, of course, is the full text of the paragraph, which is entered with the attribute of ElementID. If there is no paragraph (e.g. if the heading has all the elements or if the taxon is new) then the other elements within the AcceptedOrValidTaxonNameType should still be entered, where the data are available. The rationale for this is that ultimately the taxon treatments will be independently accessible through the web, to be associated in different contexts from the original BCA; because of this the full data must be included with them, whether it is so written in the original publication or not. The information, however, should be marked up using macros developed for the purpose. The elements not in the TaxonHeadingType are the PrimaryCitation and OtherCitations, both of which use the CitationType element (Fig. 16).

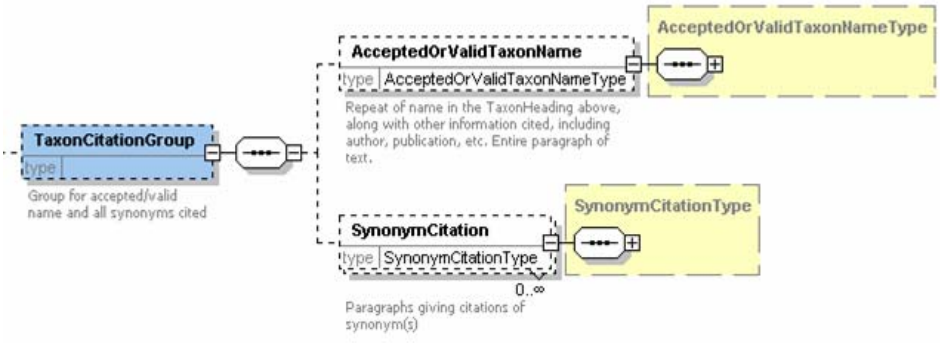


Fig. 14. The TaxonCitation Group, showing the two component Types.

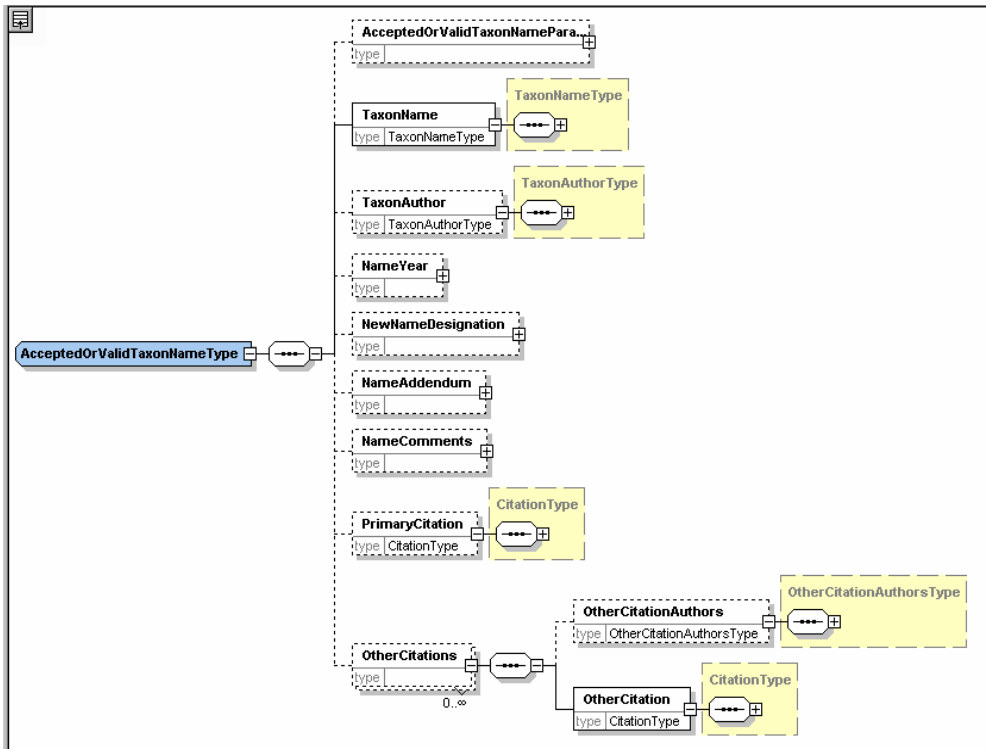


Fig. 15. AcceptedOrValidTaxonNameType

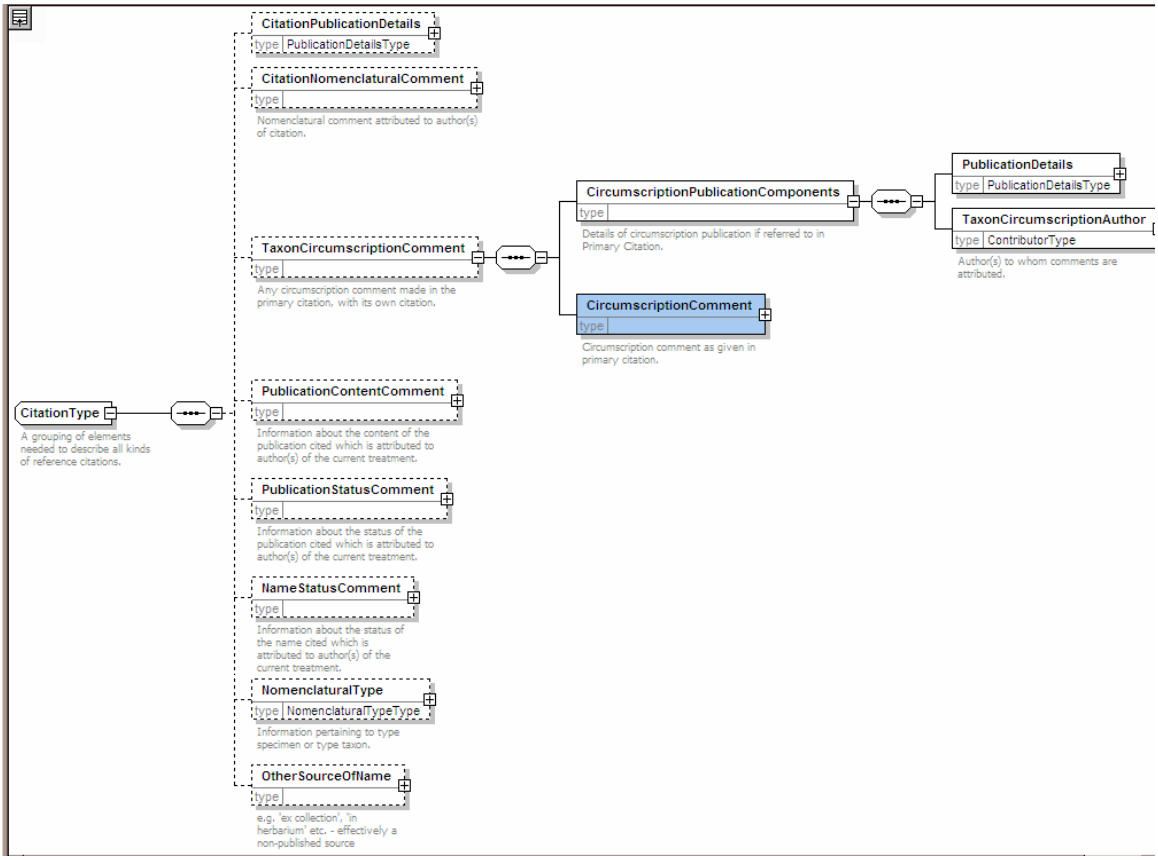


Fig. 16. CitationType complex element. The elements holding text strings have been omitted.

The CitationType is one of the most complex container elements, and includes several other 'Types'. It is designed as a grouping of elements needed to describe name citations, and consequently makes allowance for a wide variety of possible constituents. Some of the elements are simple; the CitationNomenclaturalComment (for any nomenclatural comment attributed to author(s) of citation), the PublicationContentComment (information about the content of the publication cited which is attributed to author(s) of the current treatment), the PublicationStatusComment (information about the status of the publication cited which is attributed to author(s) of the current treatment), the NameStatusComment (information about the status of the name cited which is attributed to author(s) of the current treatment) and the OtherSourceOfName (e.g. 'ex collection', 'in herbarium' etc. - effectively a non-published source) are all straightforward, and only hold a text string. The TaxonCircumscriptionComment (any circumscription comment made in the primary citation, with its own citation) is rather more complex, including as it does the comments made, together with the publication details of their source and of course their author(s). The latter element is made up as are other publication author complex elements in the schema. The TaxonCircumscriptionComment includes a PublicationDetailsType element (Fig. 17), also found containing the CitationPublicationDetails.

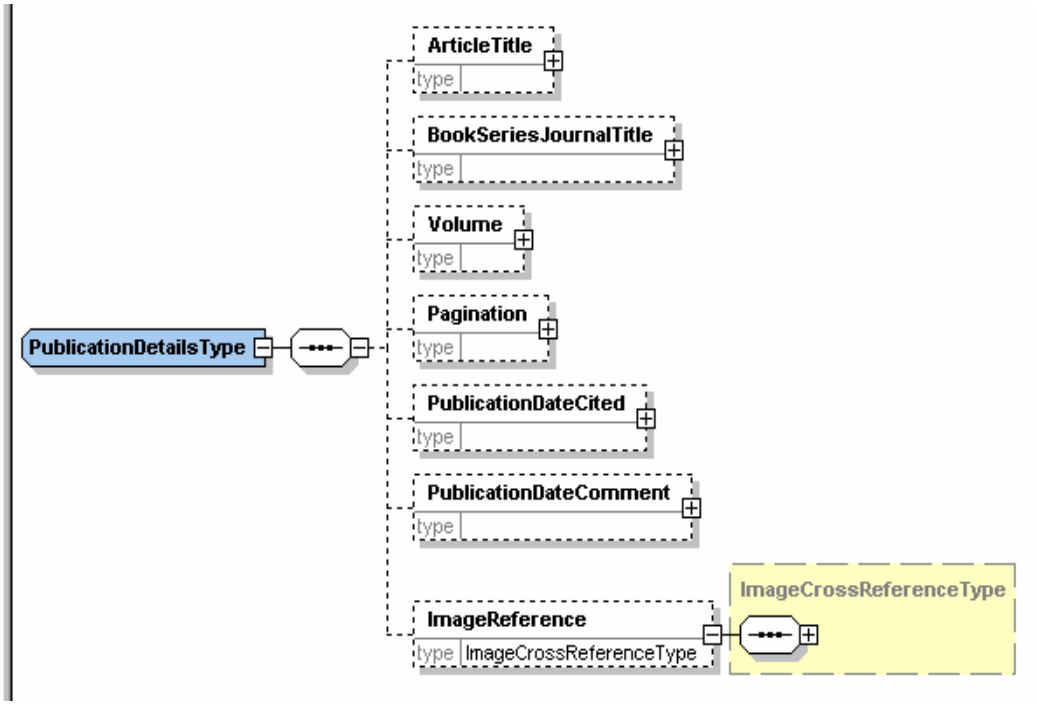


Fig. 17. PublicationDetailsType. The elements holding text strings have been omitted.

The PublicationDetailsType element contains the elements needed to describe all kinds of cited publications. With the exception of the ImageCrossReference element (Fig. 34), it contains only straightforward elements that simply act as a repository for text strings, and all of which have the attribute 'Explicit'. The PublicationDateCited refers to the date of publication given in the citation, and should any comment on the cited date have been made by the treatment author or other cited person (including and alternative publication date, if different from the cited date), this is placed in the PublicationDateComment element.

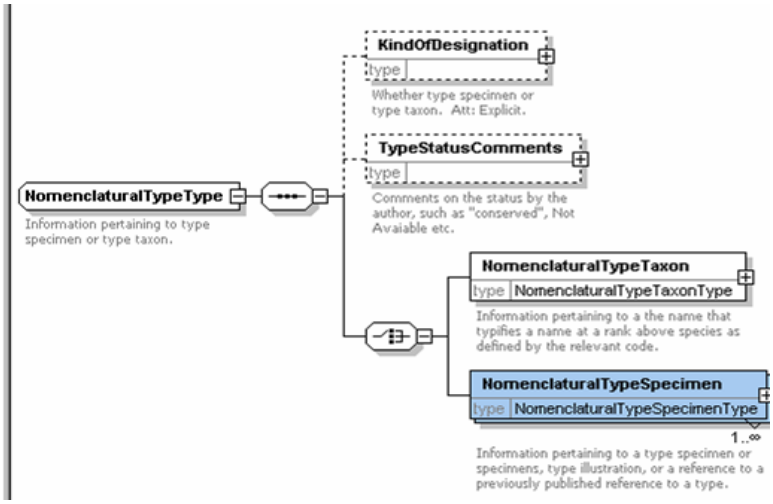


Fig. 18. NomenclaturalTypeType. The elements holding text strings have been omitted.

The AcceptedOrValidName Paragraph frequently contains a type designation or statement, and this is accommodated in the NomenclaturalType complex element (Fig. 18), which holds information pertaining to a type specimen or a type taxon. Aside from two simple text strings, the KindOfDesignation (i.e. whether the entry refers to a type specimen or type taxon) and the TypeStatusComments (comments on the status by the author, such as 'conserved', 'Not Available' etc.), the Type contains a choice between a NomenclaturalTypeTaxon (information pertaining to a the name that typifies a name at a rank above species as defined by the relevant code) and a NomenclaturalTypeSpecimen (information pertaining to a type specimen or specimens, type illustration, or a reference to a previously published reference to a type).

The NomenclaturalTypeTaxon (Fig. 19) has a fairly straightforward set of holders for the name elements, since it is unlikely that any type would be cited below the species level. The NomenclaturalTypeTaxonAuthors element is the same composition as the TaxonAuthors element. The other component, the NomenclaturalTypePublication, is of the PublicationDetailsType already discussed (Fig. 17).

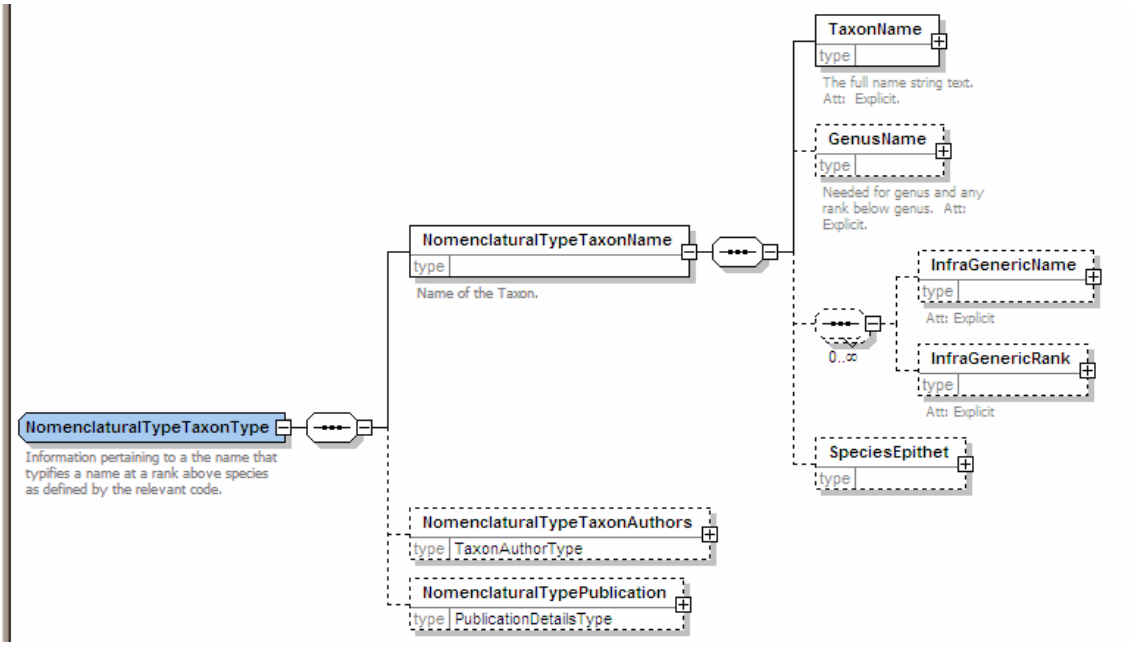


Fig. 19. NomenclaturalTypeTaxonType. The elements holding text strings have been omitted.

The NomenclaturalTypeSpecimenType (Fig. 20) also contains the PublicationDetailsType for the TypeCitation, and also the author details for the TypeCitation, where there is information pertaining to a type which is an illustration or a reference to a previously published work. The TypeStatus is a simple string containing any statement from the text on the type designation applying to either specimens or taxa (e.g. monotypy, holotype, iconotype, combined description, Lectotype [in secondary citations] etc.).

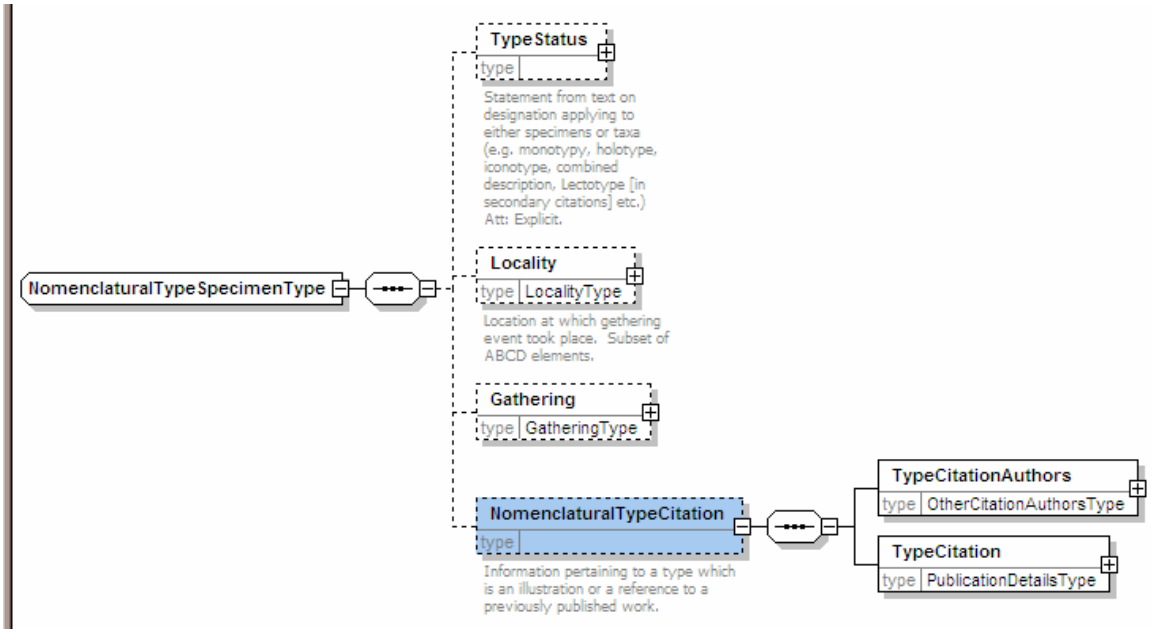


Fig. 20. NomenclaturalTypeSpecimenType. The elements holding text strings have been omitted.

The Locality and Gathering elements, which together are similar to the 'Gathering' element of ABCD, are treated separately. The Locality element (cf 'GatheringSite' of ABCD) contains the locality data, separated into 'levels' (Fig. 21). Of these 'Level0' (= 'ContinentOrOcean') will include data at the level of continent or ocean, 'Level1' (cf. 'Country') political data at the level of Country or equivalent, 'Level2' (cf 'NamedAreas') political data at the first level below Country, e.g., state, province, etc., and 'Level3' (also cf 'NamedAreas') political data at the second level below Country, e.g. county, province, district, etc. The DetailedLocation element will include any cited location other than the above, and is similar to the 'NearNamedPlace' element of ABCD, although not broken down into the place and the 'NearNamedPlaceRelationTo' elements. Altitude and Depth cover the same data as the elements of the same name in ABCD, but again are not broken down into subelements in the same way, comprising the units used, the figure (or range of figures) and any qualification, such as 'approximately' etc. The TIL will be used to make these data interoperable with other datasets. The Georeference element (Fig. 22) includes the full text string, but also a choice between the two most common types of georeference (Latitude and Longitude and Decimal Latitude and Decimal Longitude), and any other type of georeference that might be used. ABCD employs several alternatives, and the schemas might be mapped to one another through the TIL.

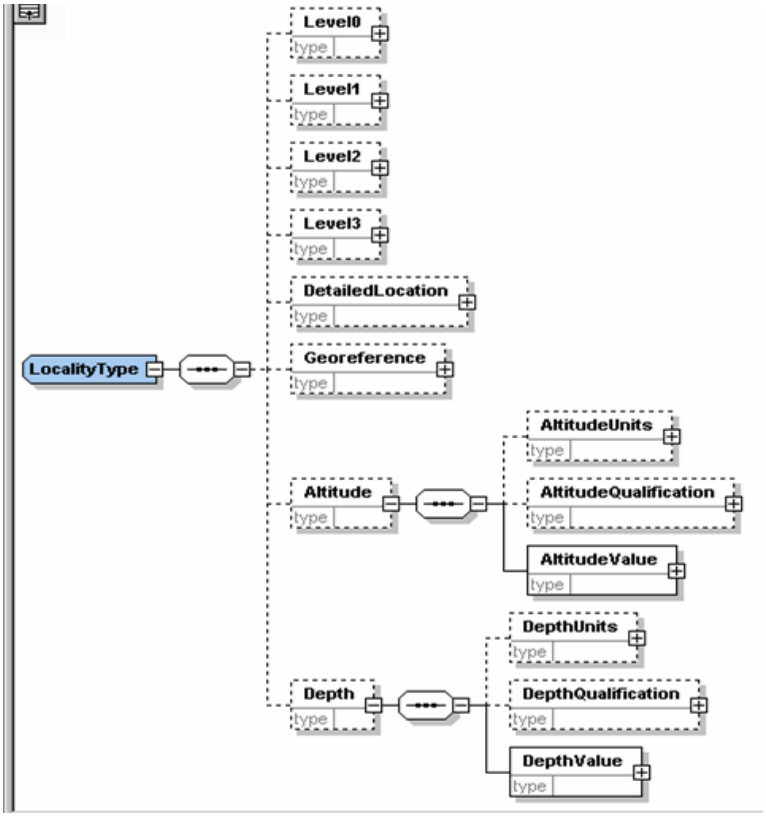


Fig. 21. LocalityType complex element. The elements holding text strings have been omitted. The expanded Georeference element is given in Fig. 22.

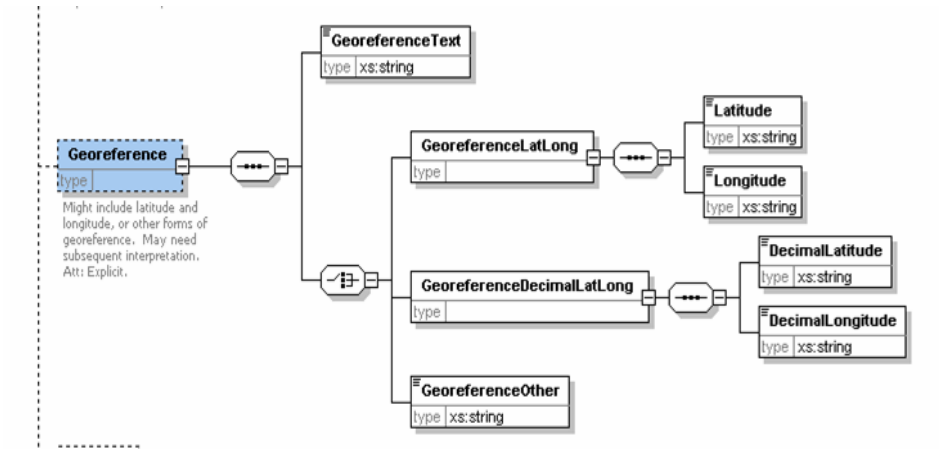


Fig. 22. Georeference complex element.

The GatheringType (Fig. 23) is similar to the ABCD GatheringType (in part), although the component elements are of simpler construction and do not allow the fine detail possible in ABCD. However, there is sufficient congruence to allow mapping within the TIL. Much of the detail in the ABCD schema, if present in taxonomic text, might be accommodated as a simple text string in CommentsAboutEntireGathering.

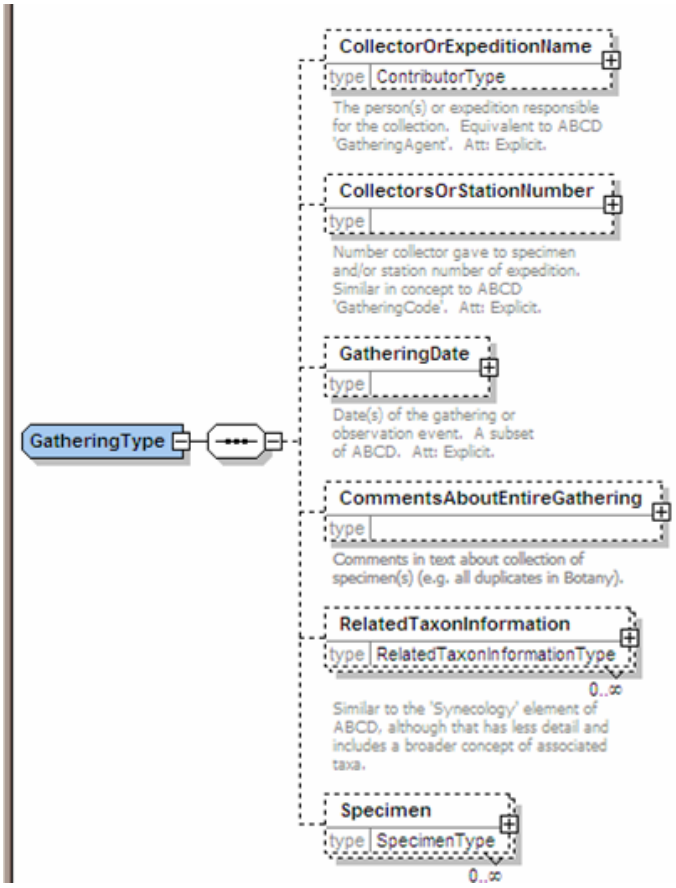
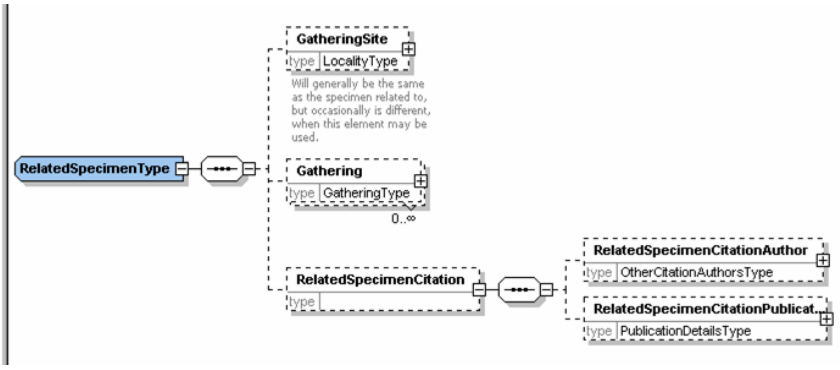


Fig. 23. GatheringType complex element. The elements holding text strings have been omitted.

The RelatedSpecimenType (Fig. 24) is made up of a LocalityType and a GatheringType (since in some circumstances there might be different information attached to related specimens than to the original taxon). Because any specimen might have been recorded from the literature rather than an original specimen, a RelatedElementCitation element has also been included, comprised of an OtherCitationAuthorsType and a PublicationDetailsType.

The RelatedTaxonInformationType (Fig. 25) deals with any taxon that is associated with the taxon in question, such as host, parasite, symbiont etc. It is similar to the 'Synecology' element of ABCD, although that has less detail and includes a broader concept of associated taxa. Most of the contained elements simply hold a text string. One exception is the RelatedTaxonName, but that is simply made up of a TaxonNameType and a TaxonAuthorType. The related taxon might be associated with specimen information, or published information.

Fig. 24. RelatedSpecimen Type



To hold citation information about the related taxon the RelatedTaxonCitedIn element comprises an OtherCitationAuthorsType and a PublicationDetailsType.

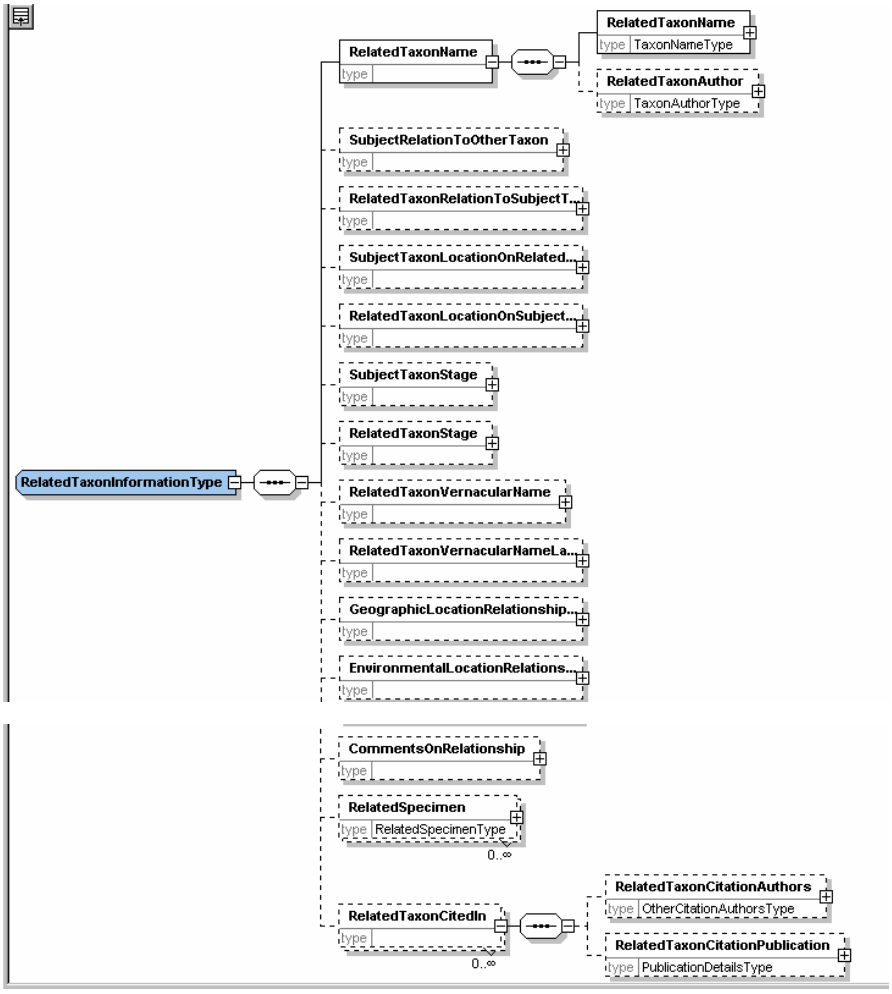


Fig. 25. RelatedTaxon InformationType. The elements holding text strings have been omitted.

The SynonymCitation Type (Fig. 26) is very similar to the AcceptedOr ValidNameType. The paragraph text is given as a single string, with an ElementID. Because synonyms may be listed in separate paragraphs or in a single paragraph, it has been structured to accommodate this, by allowing repeats of either the entire SynonymCitation or the SynonymSubsets within the SynonymCitation. The sub-elements of the 'SynonymSubsets'

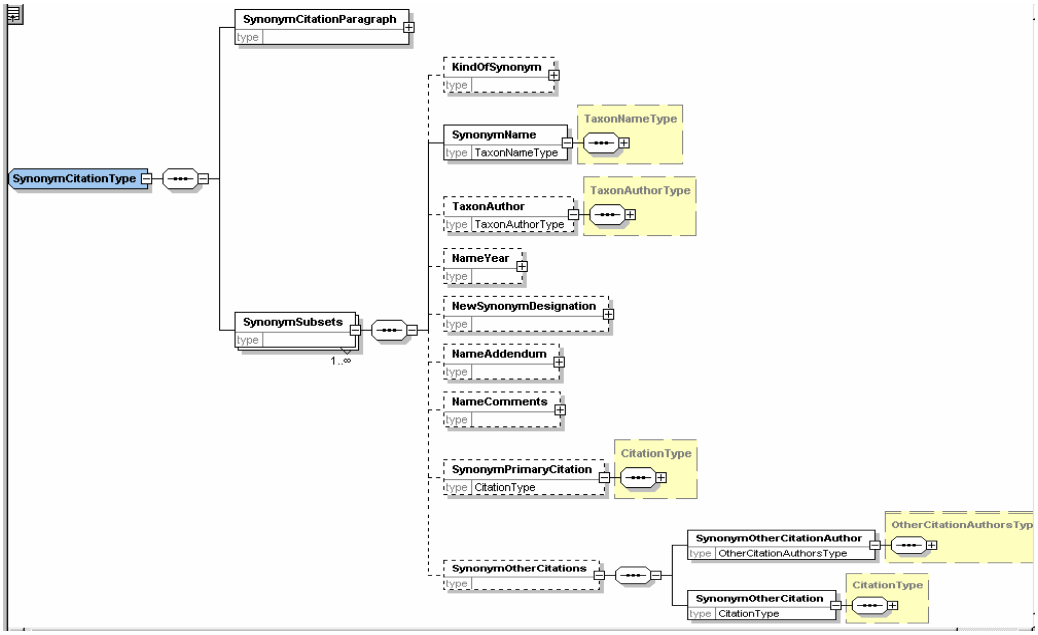


Fig. 26. SynonymCitationType. The elements holding text strings have been omitted.

are essentially the same as in the AcceptedOrValidNameType, with the addition of the KindOfSynonym element, which might contain a statement that the name is for example, a Basionym of Accepted Name, Original Name of Accepted Name, Synonym, Pro Parte Synonym etc.

The next element to be considered is the KeyToTaxa (Fig. 27).

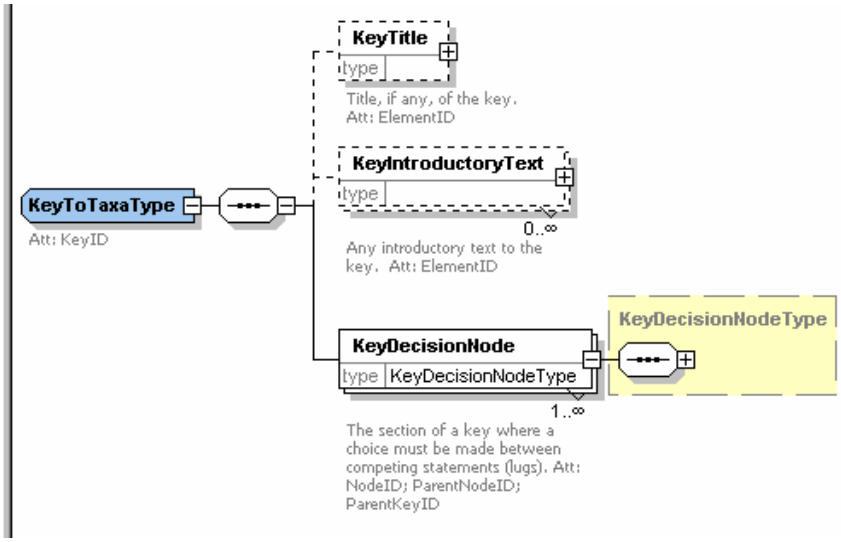


Fig. 27. KeyToTaxaType. The elements holding text strings have been omitted.

The KeyToTaxa type holds whatever keys are in the treatment. It is designed to hold both dichotomous and polytomous keys, and will accommodate a variety of formats, including keys that are spread through a treatment. The KeyTitle and KeyIntroductoryText are self-explanatory. The key itself is stored in the unbounded KeyDecisionNodeType (Fig. 28). In the BCA the couplets and lugs of the keys are often separated though the text; the provision of both ElementIDs and IDREFs within the key allows the components to be both placed in the original position within the text, but also reconstituted as a stand-alone key. The Key itself has a KeyID, and KeyDecisionNodes have a ParentKeyID to enable their placement within the appropriate key. They also have a NodeID and a ParentNodeID, to enable reconstruction of the structure of the key. The KeyDecisionNode is unbounded and recursive to allow daughter nodes. Similarly, the KeyLug has a LugID, and a ParentLugID, to enable reconstruction through the use of IDREFs. The KeyLug element comprises the full string (with an ElementID); the DecisionNodeIdentifier (which may be used to connect elements of a decision node (i.e. couplet in a dichotomous key), and is likely to be a letter or a number, which may or may not be visible with each lug); the LugIdentifier; the Characters and the KeyReference. Characters are not itemised, but simply are stated as a text string. The KeyReference may be one or more of a FollowingNumber, Taxon or KeyDecisionNode.

The DescriptionsType contains both Diagnoses and Descriptions, and each of these both in Latin and the publication language, to accommodate the most common alternatives (Fig. 29). The format of three of the four elements contained is the same, allowing for both title and text. However, in the SameLanguageDescriptions, because these are often broken down into subelements with their own heading, the structure of the contained elements is a little different.

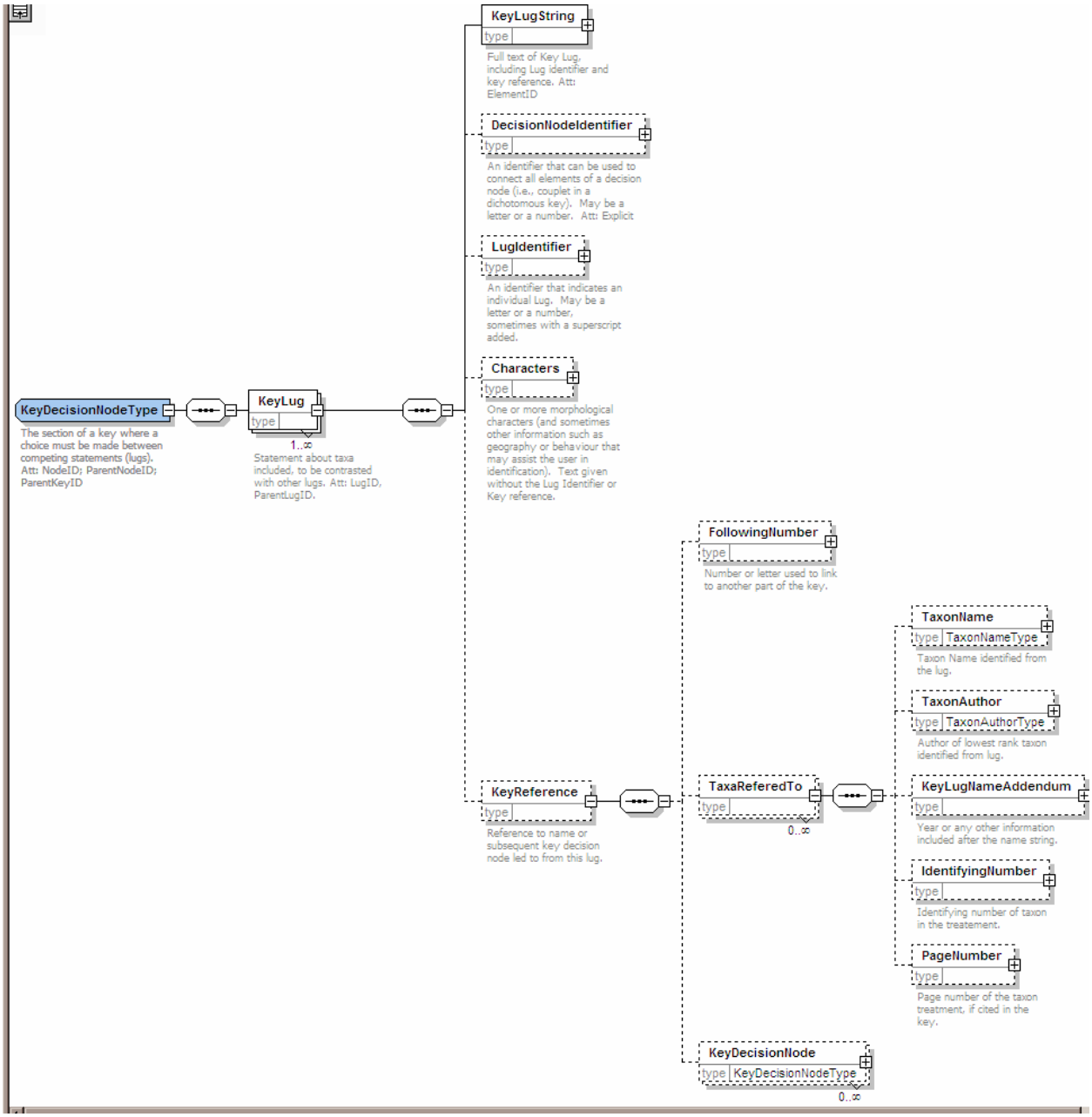


Fig. 28. KeyDecisionNodeType. The elements holding text strings have been omitted.

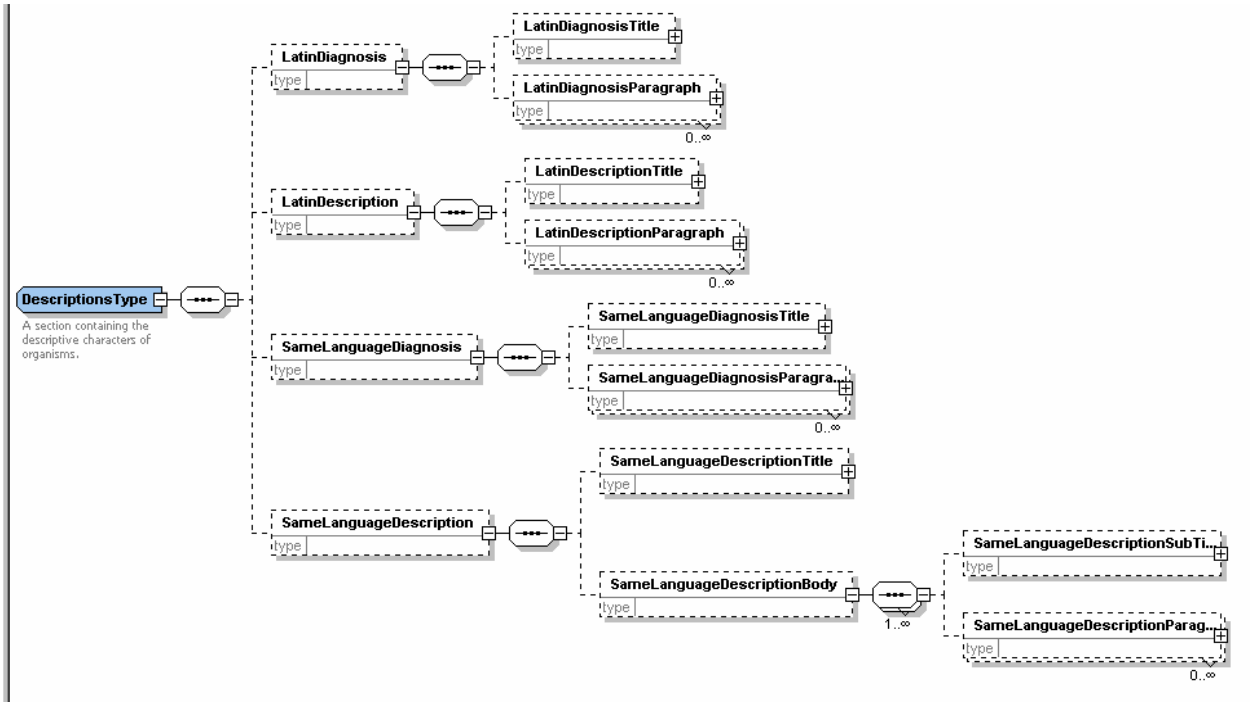


Fig. 29. DescriptionsType. The elements holding text strings have been omitted.

Within the DiscussionsType, the same format is almost always used as for the SameLanguageDescriptions (Fig. 30), and for the same reason. There are a number of Discussion Elements broken out for use where appropriate: GeneralDiscussion, MorphologyDiscussion, MaterialExaminedDiscussion, CircumscriptionDiscussion, ComparisonBetweenTaxaDiscussion, GeographicDistributionDiscussion, ConservationStatusDiscussion, BiologyAndEcologyDiscussion, TypificationDiscussion, EconomicImportanceOrUsesDiscussion, RelatedTaxaDiscussion, DerivationOfNameDiscussion, ReferencesToOtherWorksDiscussion, MolecularInformationDiscussion and ClassificationDiscussion. Of these only one, the RelatedTaxaDiscussion (Fig. 31), has a different construction. This is to allow the inclusion of more detailed information through the RelatedTaxonInformationType Element, as well as the more usual DiscussionParagraphType.

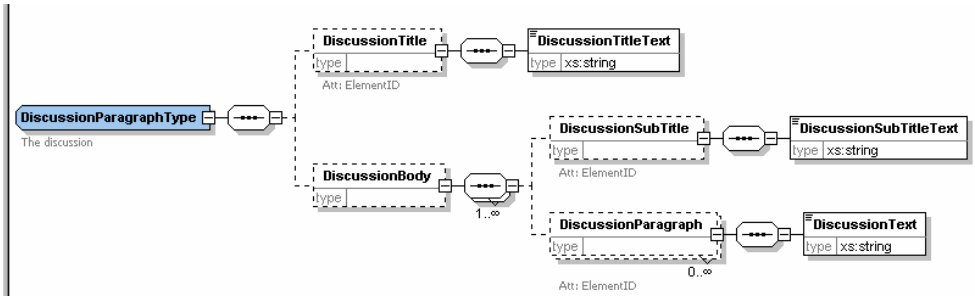


Fig. 30. DiscussionParagraphType.

The DistributionAndOrSpecimenType (Fig. 32) is straightforward, and makes use of other Types already discussed.

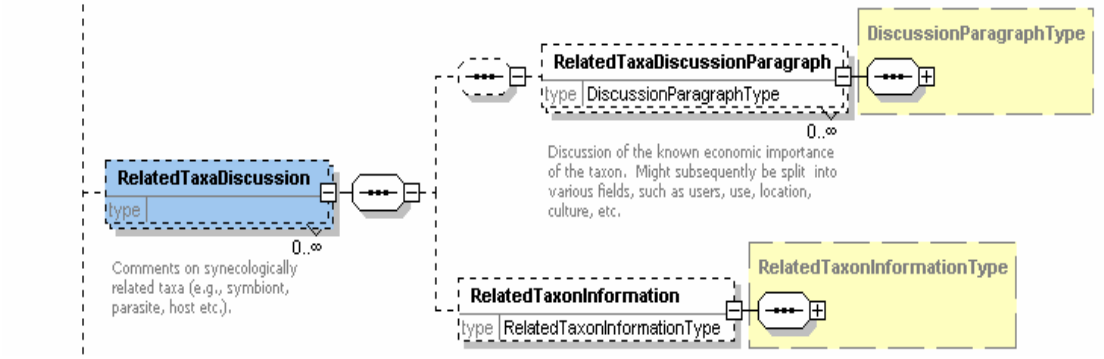


Fig. 31. RelatedTaxonDiscussion.

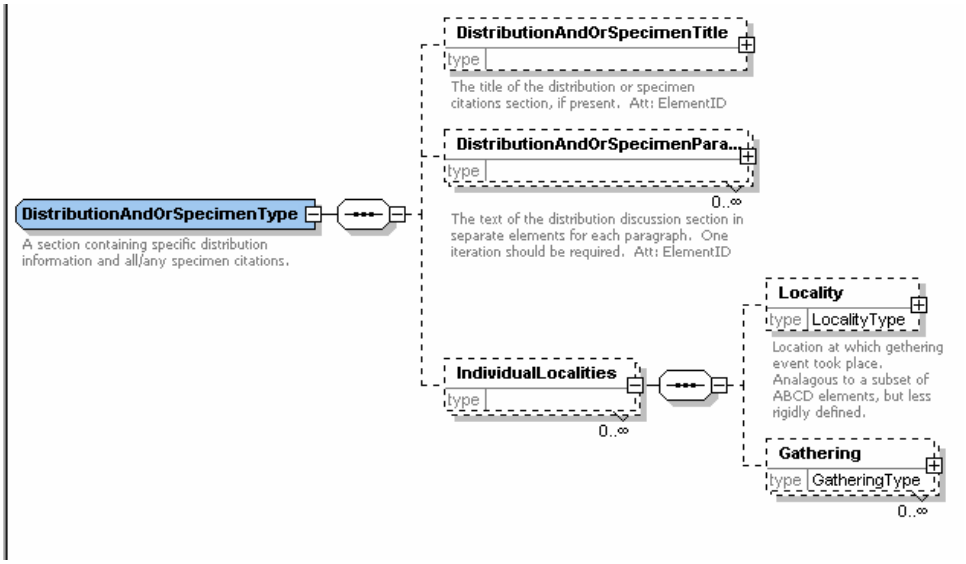


Fig. 32. DistributionAndOrSpecimenType. The elements holding text strings have been omitted.

Fig. 33. VernacularNameType. The elements holding text strings have been omitted.

The VernacularNameType (Fig. 33) allows entry of any vernacular name(s) that might be in the text, together with additional information. The VernacularName element is unbounded, to accommodate multiple names, if present.

Images and references to images are likely to be present throughout a treatment. In order to link to images from points in the text the ImageCrossReferenceType element has been included (Fig. 34). This contains the text string of the reference as in the text (e.g. 'Fig. 000'), and an IDREF to link to the image itself.

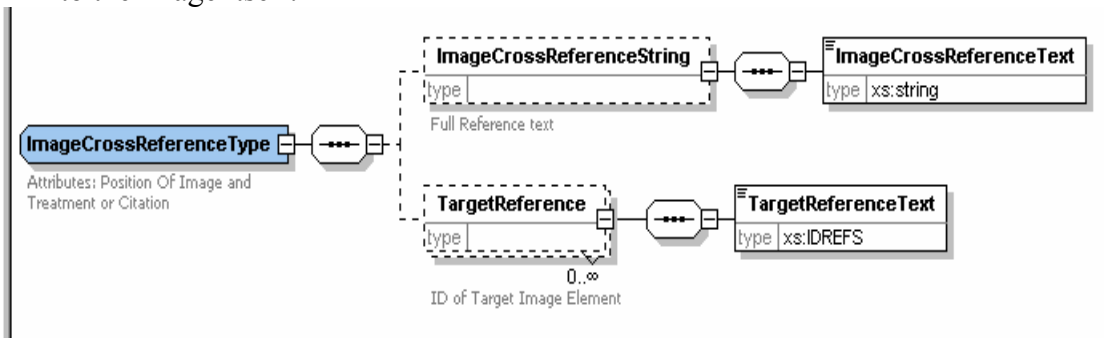


Fig. 34. ImageCrossReferenceType.

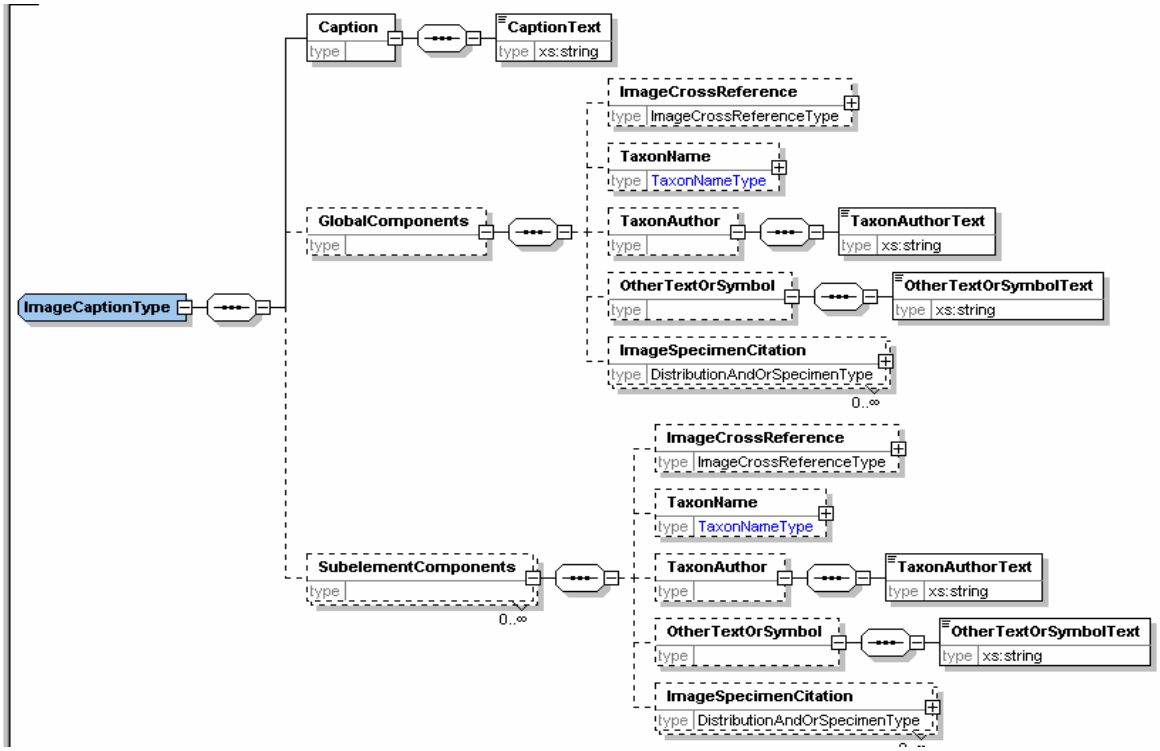


Fig. 35. ImageCaptionType

The ImageCaptionType (Fig. 35) may contain either a single set of information or refer to several taxa or specimens. Accordingly the schema allows for a single caption or an unbounded set of subelements, with the same set of elements contained. In this context the TaxonAuthor is not broken down as it is elsewhere in the schema, but recorded only as a simple string.