GeM Project Report
University of Stirling
University of Bremen

http://purl.org/net/gem
e-mail: henschel@uni-bremen.de

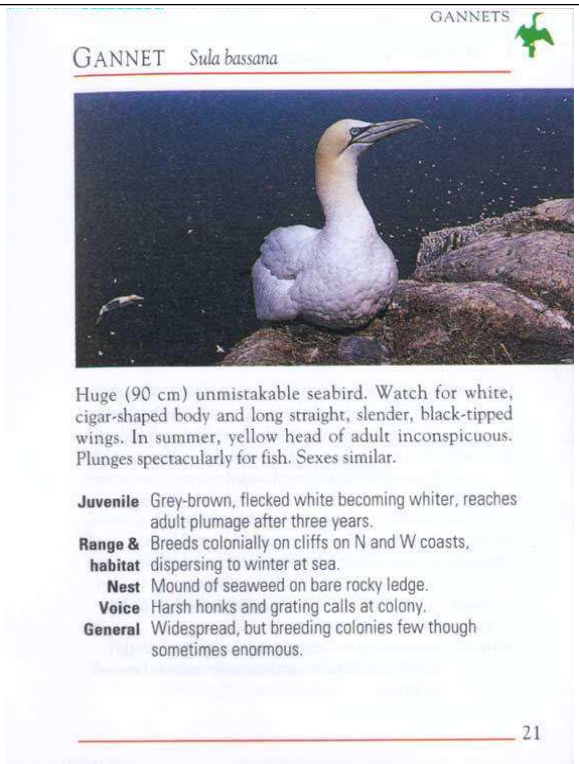
GeM Annotation: One complete example – the Flegg page

Renate Henschel

March 13, 2003

1 One complete example – the Flegg page

Here we present the complete XML annotation for one page of a bird guide following the GeM annotation scheme. The particular page used is page 21 of **Flegg, J.** (1999) Birds. Collins Gem. Italy: Harper Collins. We call it the **flegg page**; it is shown in Figure 1.


GANNETS

GANNET *Sula bassana*

Huge (90 cm) unmistakable seabird. Watch for white, cigar-shaped body and long straight, slender, black-tipped wings. In summer, yellow head of adult inconspicuous. Plunges spectacularly for fish. Sexes similar.

Juvenile Grey-brown, flecked white becoming whiter, reaches adult plumage after three years.
Range & habitat Breeds colonially on cliffs on N and W coasts, dispersing to winter at sea.
Nest Mound of seaweed on bare rocky ledge.
Voice Harsh honks and grating calls at colony.
General Widespread, but breeding colonies few though sometimes enormous.

21

Figure 1: A starting point for annotation: the flegg page

We assume that the overall structure of the GeM annotation scheme, as well as the general strategies to be employed in preparing an annotation, are known; they are described in

the annotation manual. Here we turn straight to the concrete task of moving from page to annotation. The resulting XML specifications of the annotation layers described are presented as appendices; it will be useful to turn to these during reading so as to relate concretely the description of the annotation process with the encoding that this gives.

1.1 The GeM base

Breaking out the base units of the flegg page gives us altogether 23 base units. These are: the green icon and the GANNET word at the top right, the two parts of the title (GANNET, Sula bassana), the photo, five sentences, 10 table cells (5 rows with two cells each), the page number, and two horizontal, red lines.

Label	Unit
u-21.01	GANNETS
u-21.02	(gannet icon)
u-21.03	GANNET
u-21.04	Sula Bassana
u-21.05	(horizontal red line)
u-21.06	(photo)
u-21.07	Huge (90cm) unmistakable seabird.
u-21.08	Watch for white, cigar-shaped body and long straight, slender, black-tipped wings.
u-21.09	In summer, yellow head of adult inconspicuous.
u-21.10	Plunges spectacularly for fish.
u-21.11	Sexes similar.
u-21.12	Juvenile
u-21.13	Grey-brown, flecked becoming whiter, adult plumage after three years.
u-21.14	Range and habitat
u-21.15	Breeds colonially on cliffs on N and W coasts, dispersing to winter at sea.
u-21.16	Nest
u-21.17	Mound of seaweed on bare rocky ledge.
u-21.18	Voice
u-21.19	Harsh honks and grating calls at colony.
u-21.20	General
u-21.21	Widespread, but breeding colonies few though sometimes enormous.
u-21.22	(horizontal red line)
u-21.23	21

Table 1: Labelling the base units

We label these following the numbering scheme suggested in the annotation manual: i.e., a sequence of labels each beginning with the "u-21.", indicative of the fact that we are concerned with page 21 of the document from which the page was taken. The units are then in general numbered from top to bottom, although this can be freely deviated from if inconvenient. No information apart from convenient labelling is hidden in the numbering. For the flegg page, we have the list shown in Table 1.

1.2 The layout structure

Segmentation. From the determined 23 base units, the text of the main paragraph, consisting of the five sentences (u-21.07 until u-21.11), form together a single unit as far as the layout is concerned. The remaining 18 base units ????? each function as separate layout units. So we have altogether 19 layout units. These are also labelled following the strategy set out on the manual; we incorporate where possible the numbering of the base units, appending these to the prefix "lay-21." for the text elements and, for mnemonic purposes, some identifying label such as "line-21." or "photo-21" for non-text elements. The precise labelling and naming of the units is again completely arbitrary—the meaning is given solely by their relationships with the other layers of annotation.

Following the annotation scheme, each layout unit has a unique identifier (e.g., "lay-21.03") and a reference to the base unit or base units to which it corresponds (e.g., for "lay-21.03" the base unit "u-21.03"). The XML representation for one layout-unit would then look e.g. as follows:

```
<layout-unit id="lay-21.03" xref="u-21.03"/>
```

Realization. For each layout unit, we then need to code its particular typographical and visual realization. From the 19 layout units, we can differentiate 4 graphical elements: the green icon (lay-21.02), the photo (flegg-photo) and two red lines (line-21.1, line-21.2). The red lines and the icon are two-dimensional graphical elements (type="two-d-element"). The photo is of type="photo". For the red lines, we also mark two-d-element-type="line", element-style="solid", and element-width="bold". The precise attributes that are to be determined for each type of element are given in the manual. For photos and illustrations, we choose additionally between the color values black-white or color. In this case it would be colors="color". The example annotation for the photo and the lines is therefore:

```
<graphics xref="flegg-photo" type="photo" colors="color" height="5.6cm"/>
<graphics xref="line-21.1 line-21.2" type="two-d-element" two-d-element-type="line"
            element-style="solid" element-weight="bold" colors="red"/>
```

Note that this annotation does not have a separate closing tag but rather is self-contained, as indicated by the closing "/>" bracket.

All the other layout units are textually realized elements. For these we identify their particular typographical properties, drawing out commonalities in appearance by defining text elements. Thus, for example, the table cells of the first column share an identical typographical realization, as do the cells of the second column. So we need only two text elements to annotate their typographical properties. Furthermore, the typographical realization of flegg-text is entirely different from that of the table columns. This then requires an extra text element for flegg-text.

The other text elements all differ from each other, and so receive their own text element annotation.

It is not quite clear in this page, whether the page number shares its typographical layout with flegg-text. We have chosen to annotate a slightly smaller font-size, so we write for the

page number also an own text element. This makes altogether 7 text elements (GANNETS, GANNET, Sula basana, flegg-text, column 1 of the table, column 2 of the table, page number).

The XML annotation for the typographical features of the first column of the flegg-table would be, for example:

```
<text xref="lay-21.12 lay-21.14 lay-21.16 lay-21.18 lay-21.20"
      font-family="sans-serif"
      font-size="10" font-style="normal" font-weight="bold"
      case="mixed" justification="right" color="black"/>
```

Area model. The flegg page is structured in two columns: a big one which nearly includes everything on the page, and a small one, which is more or less a right margin which is used for some layout elements. The flegg-photo ranges into this column and the icon is situated in it. Beside this two-column structure, the flegg-page also has five rows: the area above the first red line , the area for the photo, the area for flegg-text, the area for the table, the area with the second red line and the page number. We will call this grid structure of the flegg-page "flegg-page-frame". The 2 columns and 5 rows altogether define 10 rectangular sub-areas (cell-11, cell-12, cell-21, ... cell-52) on the flegg-page.

The table in cell-41 has a more detailed structure, so we define for cell-41 a sub-area "table-frame": a table with five rows and two columns. Note that in this table the elements of the first column are right-aligned!

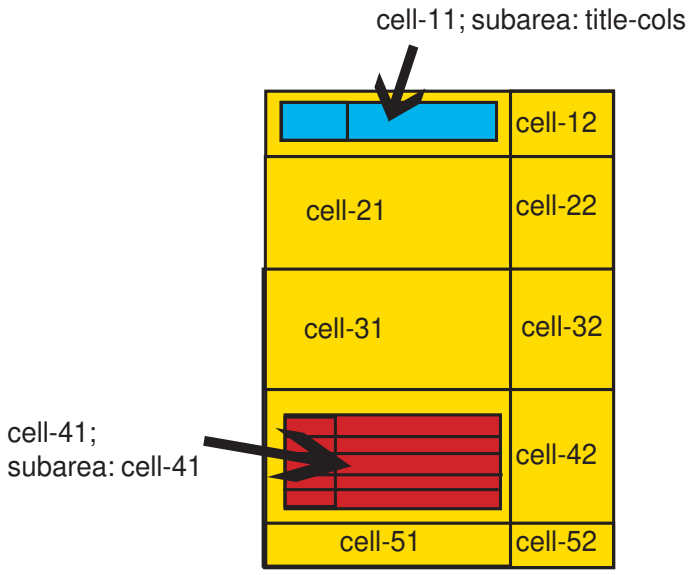


Figure 2: Area model for the flegg page showing subareas

The title of the page "GANNET Sula bassana" cannot be seen as continuous text; it has an unusually large gap between the English and the Latin name. We can describe this gap by defining another sub-area for cell-11, where the title is located. If we consider cell-11 to consist of two columns, the first about 20%, the second about 80% of the width of cell-11, then "GANNET" would sit in the first column, and "Sula bassana" in the second.

Summarizing, we have an area model for the flegg-page with 5 rows and 2 columns. Two of its sub-areas – cell-11 and cell-41 – are further partitioned into smaller areas, cell-11 into

two columns, cell-41 into a table with 5 rows and 2 columns. The page areas defined by this model are sufficient to allocate a precise location to each of the flegg pages layout units. This area model is shown graphically in Figure 2.

The relative sizes of the cells of any particular row or column are then indicated by percentages as suggested above. For example, for the case of the title area, we have the following annotation:

```
<sub-area id="title-cols" location="cell-11" cols="2" rows="1"
          hspacing="20 80" vspacing="100"/>
```

The structure of the complete area model corresponding to Figure 2 is shown in Appendix 2.

Layout structure. Looking at the flegg page, we can differentiate header and footer material and the main body of the page. The header material is the actual title (GANNET Sula bassana) on the one hand and the running head at the top right of the page (GANNETS + bird icon). The footer is the page number. The page body consists of three layout chunks: the flegg-photo, the flegg-text, and the flegg-table. The flegg-table itself has 10 layout units, the actual text in the individual table cells. The two red lines serve to recognize the separation between header and footer material, and the main page body. We view them therefor as layout-children of the flegg-page. The other, here not followed possibility would be to perceive them as part of the page body. The resulting layout structure is then as shown in Figure 3.

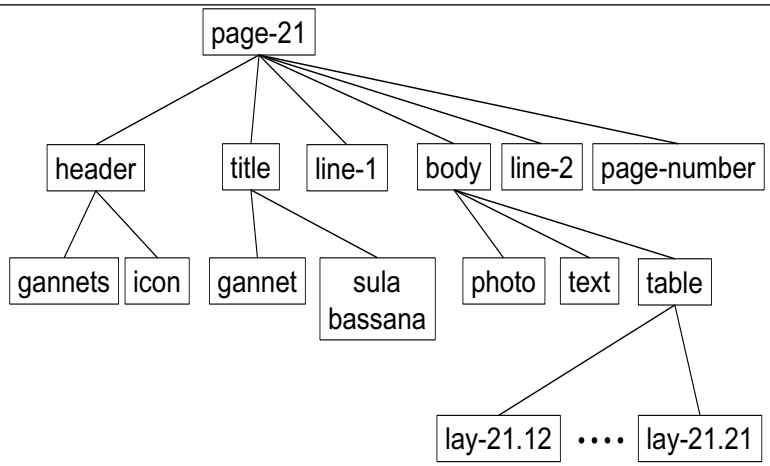


Figure 3: Layout structure for the flegg page

In order to determine precisely the place that a certain layout leaf/chunk occupies on the page, we use the area model. We add to every layout chunk and leaf a location value and an area-ref value: this then serves to locate any element of the layout within an area specified by the area model. For example, the flegg-photo occupies the entire row-2 of the flegg-page-frame; flegg-text is located in cell-31 of the flegg-page-frame, and flegg-table in cell-41 of the flegg-page-frame. Cell-41, furthermore, has its own internal area model, which allows the precise location of the descendent layout units also. Thus: flegg-table's children (the layout-units lay-21.12 ... lay-21.21) are all in cell-41 of the flegg-page-frame, but at different sub-areas. To give them a precise location, we use the sub-area "table-frame" and allocate them values with respect to this frame. This means that, e.g., "Nest" is located in cell-31 of the table-frame. All the location values hold with respect to a particular frame drawn from the area model. Which frame we mean is always marked under the attribute "area-ref".

An example annotation is therefore:

```
<layout-chunk id="title" location="cell-11" area-ref="flegg-page-frame">
```

This says that the layout chunk "title" is located in the first column of the first row of the grid area identified as "flegg-page-frame".

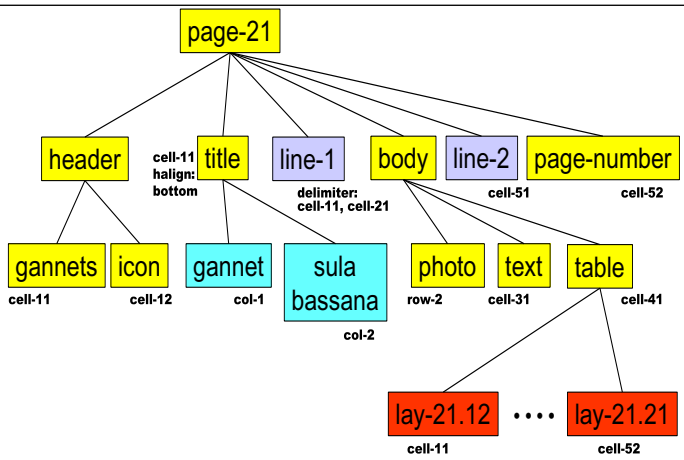


Figure 4: Layout structure for the flegg page with locations

The fact that the labels of the flegg-table (lay-21.12, lay-21.14, lay-21.16, lay-21.18, lay-21.20), which are located in column 1 of the table-frame, are aligned at the right side of this column has to be annotated by adding **valign**="right" for these five layout-leafs. Another non-default alignment is to be found with the title in cell-11 of the page-frame. Both layout-leafs (GANNET, Sula bassana) are placed at the bottom of this cell. So these two layout-leafs get marked with **halign**="bottom". Also the GANNETS word beneath the icon is not placed corresponding to the default (at the left top edge of cell-11). Instead we find it at the right top edge.

1.3 The rhetorical structure

Segmentation. The flegg page has 9 RST segments, and 8 mini-segments. First of all, sentences (even elliptical ones) are marked as segments. We have here four sentences in the text below the photo:

> *Huge (90cm) unmistakable seabird. Watch for white, cigar-shaped body and long straight, slender, black-tipped wings. In summer, yellow head of adult inconspicuous. Plunges spectacularly for fish. Sexes similar.*

The XML specification for one segment has an id- and an xref-value. The xref value consists of one or more base unit ids, which form this segment.[1] The id value is the identifiying label for this segment, and usually starts with the prefix "s-". The following is an example XML specification for a segment:

[1]Sentences which are disrupted into two base units by page or column breaks are examples for segments consisitng of more than one base unit.

```
<segment id="s-21.08" xref="u-21.08">Watch for white, cigar-shaped body and
        long straight, slender, black-tipped wings.</segment>
```

In addition to these easily recognizable segments, we have other material on the page which contributes to the semantic content: the table consisting of five rows, the title line (GANNET + Sula bassana), the photo. As next step, we analyze the table rows, in particular the semantic relation which holds between the content of the first and the content of the second cell in each row. We can find the following relations:

(row-1) property-ascription: `grey-brown(juvenile-gannet)`
(row-2) about relation: cell-22 is about "habitat"
(row-3) class-ascription: `ISA(nest, seaweed-mound)`
(row-4) class-ascription: `ISA(voice, honks)`
(row-5) projection: It can generally be said that (The gannet is) Widespread, but breeding colonies (are) few though sometimes enormous.

Analyzing the table rows, we realize that three of them represent abbreviations for a simple *is*-clause:

> (A/the) Juvenile (gannet) (is) flecked becoming whiter, reaches adult plumage after three years.
>
> (The) Nest (of a gannet) (is a) Mound of seaweed on bare rocky ledge.
>
> (The Gannet's) Voice (is characterized by) Harsh honks and grating calls at colony.

The same holds for the title line: *GANNET (is in Latin) Sula bassana*, We mark every base unit which participates in an *is/has*-clause as a mini-segment.[2] Hence we have now eight mini-segments, the cells of row-1, row-3 and row-4 of the table, and two which form the title line. The minisegment pairs representing row-1, row-3, row-4 and the title line themselves form four mini-spans. The relation which holds between the two contributing mini-segments are class-ascription (row-3, row-4), property-ascription in row-1 and the identifying relation for the title line mini-span. The XML representation for a mini-segment and a mini-span would then look e.g. as follows:

```
<mini-segment id="s-21.16" xref="u-21.16">Nest</mini-segment>
<mini-segment id="s-21.17" xref="u-21.17">Mound of seaweed on bare rocky
              ledge.</mini-segment>
<mini-span id="s-21.25" attribute="s-21.16" attribuend="s-21.17"
           relation="class-ascription"/>
```

Note that the text inside the XML tags is not a necessary part of the annotation.

There are still some base units which are not marked as RST segments or minisegments yet: row-2 and row-5 of the table, the photo, the icon, and the GANNETS beside the icon.

In table row 2, the relation between the content of the first and of the second table cell is an about-relation. That means cell-21 functions as a heading for cell-22, and is therefor a

[2]Note that is/has-clauses are normally marked as segments, not as two mini-segments. Only the ?????deliberate layout separation of the two participants of such a clause as e.g. in a table makes it a candidate for mini-segment markup.

navigational element, but not a RST segment. cell-22 stands as a segment by its own. This is also supported by the fact that cell-22 is a clause-like phrase with a finite verb, whereas cell-12, cell-32, and cell-42 are NPs only. Row-5 seems to be an abbreviated form of the sentence *It can generally be said that the gannet is widespread, but breeding colonies are few though sometimes enormous.*. If we agree with this interpretation, table cell-51 is an abbreviation of *It can generally be said*, and the table cell-52 is an abbreviation of *the gannet is widespread, but breeding colonies are few though sometimes enormous*. The relation between the cell-51 and the cell-52 text piece is **not** a be-relation. Also the text in cell-52 can rather stand by itself as a clause, and does not need the text in cell-51. This has lead us to decide that cell-51 and cell-52 form each a segment on their own right.[3]

The Gannet photo as the central piece of information on the page is marked as segment. The green icon at the top right of the page has a similar function as a running head. To realize this one ahs to know the general structure of the bird guide: All bird pages belonging to one and the same family carry an identical icon togetehr with the family name. Thus, the icon and the "GANNETS" word are navigational elements and not marked as segments.

Altogether we have now identified 9 RST segments: the five sentences in the text part (s-21.07... s-21.11), the photo (s-21.06), the table cell-22 (s-21.15), the table cell-51 and cell-52 (s-21.20, s-21.21), and eight mini-segments: the table cells in row 1,3,4 (s-21.12, s-21.13, s-21.16, s-21.17, s-21.18, s-21.19), and the title line (s-21.03, s-21.04). They form four mini-spans, which have the following internal structure in XML annotation:

```
<mini-span id="s-21.22" attribuend="s-21.03" attribute="s-21.04"
           relation="identification">GANNET  Sula bassana</mini-span>
<mini-span id="s-21.24" attribuend="s-21.12" attribute="s-21.13"
           relation="property-ascription">Juvenile Grey-brown, flecked becoming
           whiter,  adult plumage after three years.</mini-span>
<mini-span id="s-21.25" attribute="s-21.16" attribuend="s-21.17"
           relation="class-ascription">Nest Mound of seaweed on bare rocky
           ledge.</mini-span>
<mini-span id="s-21.26" attribute="s-21.18" attribuend="s-21.19"
           relation="class-ascription">Voice Harsh honks and grating calls at
           colony.</mini-span>
```

RST structure. The segments and mini-spans are the basic elements which are grouped together in the RST structure. The rhetorical structure which we propose for the page is shown in Figure 5. The mini-spans are also presented in the figure. Note that 'flegg-gannet' is the span which comprises the rhetorical content of the entire flegg page.

1.4 Navigation structure

We mark as navigation elements pointers to other pages, books etc., titles and headings, running heads and page numbers. Our example page has 7 navigational elements: the bird

[3] As a side effect, one can see from this analysis that the table on this page is not well motivated, because a table layout can only effectively support the reader's needs, if the relations which hold between column-one cells and column-two cells are rather similar.

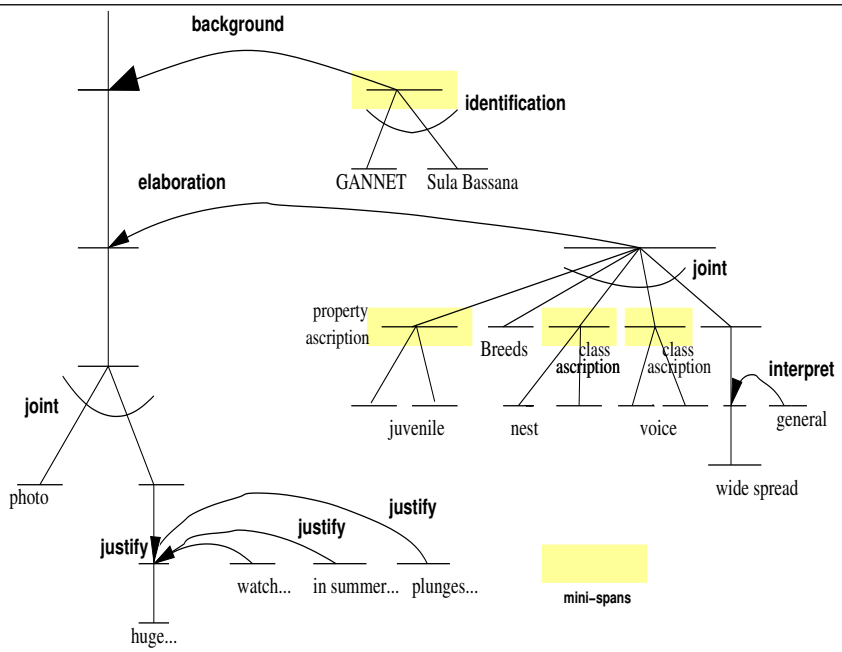


Figure 5: RST structure for the flegg page with locations

icon, the word GANNETS beside it, the title (GANNET Sula bassana), the table cell-21 (Range and habitat), the page number and the two red lines. We classify these navigation elements whether they are entries, indices or pointers. Entries are names for RST spans or segments which are usually used as heading. On the flegg page, these are the title line, the table cell-21, the icon and the word 'GANNETS': The title line is a name for the RST span 'flegg-gannet' (the entire page); the "Range and habitat" table cell is the heading for the adjacent table cell – the segment s-21.15. Also the bird icon and the word GANNETS are names of a bigger – not in the corpus included – RST span, which contains the Gannet page. The page number, on the other hand, is the name for the layout unit "flegg-page-21", and therefor an index. For the time being, we do not specify markup for the lines in the navigation base, although it is clear that these lines play a role in the readers navigation through the page. So, we have four entries and one index defined on the flegg page. There are no pointers at the flegg page.

The precise and complete annotations required can be seen in the appendices.

Appendix 1: XML file of the base units

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE gemBase SYSTEM ".../GeM/corpus/dtds/gem-base.dtd">
<gemBase>
  <unit id="u-21.01">GANNETS</unit>
  <unit id="u-21.02" alt=‘‘gannet-icon’’/>
  <unit id="u-21.03">GANNET</unit>
  <unit id="u-21.04">Sula Bassana</unit>
  <unit id="u-21.05">----------------------------------------</unit>
  <unit id="u-21.06" alt=‘‘gannet-photo’’/>
```

```
  <unit id="u-21.07">Huge (90cm) unmistakable seabird.</unit>
  <unit id="u-21.08">Watch for white, cigar-shaped body and long straight,
                   slender, black-tipped wings. </unit>
  <unit id="u-21.09">In summer, yellow head of adult inconspicuous. </unit>
  <unit id="u-21.10">Plunges spectacularly for fish.</unit>
  <unit id="u-21.11">Sexes similar. </unit>
  <unit id="u-21.12">Juvenile</unit>
  <unit id="u-21.13">
          Grey-brown, flecked becoming whiter, adult plumage after three years.
          </unit>
  <unit id="u-21.14">Range and habitat</unit>
  <unit id="u-21.15">Breeds colonially on cliffs on N and W coasts, dispersing
                    to winter at sea.</unit>
  <unit id="u-21.16">Nest</unit>
  <unit id="u-21.17">Mound of seaweed on bare rocky ledge.</unit>
  <unit id="u-21.18">Voice</unit>
  <unit id="u-21.19">Harsh honks and grating calls at colony.</unit>
  <unit id="u-21.20">General </unit>
  <unit id="u-21.21">Widespread, but breeding colonies few though sometimes enormous.
                    </unit>
  <unit id="u-21.22">----------------------------------------</unit>
  <unit id="u-21.23">21</unit>
</gemBase>
```

Appendix 2: XML file of the layout structure

```
<?xml version="1.0"?>
<!DOCTYPE gemLayout SYSTEM ".../GeM/corpus/dtds/area-gem-layout.dtd">
<gemLayout>
  <segmentation>
    <layout-unit id="lay-21.01" xref="u-21.01"/>
    <layout-unit id="lay-21.02" xref="u-21.02"/>
    <layout-unit id="lay-21.03" xref="u-21.03"/>
    <layout-unit id="lay-21.04" xref="u-21.04"/>
    <layout-unit id="line-21.1" xref="u-21.05"/>
    <layout-unit id="flegg-photo" xref="u-21.06"/>
    <layout-unit id="flegg-text" xref="u-21.07 u-21.08 u-21.09 u-21.10 u-21.11"/>
    <layout-unit id="lay-21.12" xref="u-21.12"/>
    <layout-unit id="lay-21.13" xref="u-21.13"/>
    <layout-unit id="lay-21.14" xref="u-21.14"/>
    <layout-unit id="lay-21.15" xref="u-21.15"/>
    <layout-unit id="lay-21.16" xref="u-21.16"/>
    <layout-unit id="lay-21.17" xref="u-21.17"/>
    <layout-unit id="lay-21.18" xref="u-21.18"/>
    <layout-unit id="lay-21.19" xref="u-21.19"/>
    <layout-unit id="lay-21.20" xref="u-21.20"/>
    <layout-unit id="lay-21.21" xref="u-21.21"/>
    <layout-unit id="line-21.2" xref="u-21.22"/>
    <layout-unit id="flegg-page-no" xref="u-21.23"/>
  </segmentation>
  <realization>
    <!-- textual elements with similar typographical features -->
    <text xref="lay-21.01" font-family="serif" font-size="8" font-style="normal
          font-weight="normal" case="caps" justification="no" color="black">GANNETS</text>
    <text xref="lay-21.03" font-family="serif" font-size="14" font-style="normal"
          font-weight="normal" case="caps" justification="no" color="black''>GANNET</text>
    <text xref="lay-21.04" font-family="serif" font-size="12" font-style="normal"
          font-weight="normal" case="caps" justification="no" color="black">
             Sula bassana</text>
    <text xref="flegg-text" font-family="serif" font-size="11" font-style="normal"
          font-weight="normal" case="mixed" justification="justified" color="black"/>
    <text xref="lay-21.12 lay-21.14 lay-21.16 lay-21.18 lay-21.20" font-family="sans-serif"
          font-size="10" font-style="normal" font-weight="bold"
          case="mixed" justification="right" color="black"/>
    <text xref="lay-21.13 lay-21.15 lay-21.17 lay-21.19 lay-21.21"
          font-family="sans-serif" font-size="11" font-style="normal" font-weight="normal"
          case="mixed" justification="left" color="black"/>
    <text xref="flegg-page-no" font-family="sans-serif" font-size="10" font-style="normal"
          font-weight="normal" case="mixed" justification="no" color="black">21</text>
    <!-- graphical elements -->
    <graphics xref="lay-21.02" type="two-d-element" two-d-element-type="icon" color="green"/>
    <graphics xref="flegg-photo" type="photo" color="color" height="5.6cm"/>
    <graphics xref="line-21.1 line-21.2" type="two-d-element"  two-d-element-type="line"
              element-style="solid" element-weight="bold" color="red"/>
  </realization>
```

```
  <area-model>
    <area-root id="flegg-page-frame" cols="2" rows="5" hspacing="90 10"
               vspacing="10 30 20 38 2" height="16cm" width="14cm">
      <sub-area id="title-cols" location="cell-11" cols="2" rows="1" hspacing="20 80"
                vspacing="100"/>
      <sub-area id="table-frame" location="cell-41" cols="2" rows="5" hspacing="15 85"
                vspacing="flexible"/>
    </area-root>
  </area-model>
  <layout-structure>
    <layout-root id="flegg-page-21">
      <layout-chunk id="flegg-header" location="multi" area-ref="flegg-page-frame">
        <layout-leaf xref="lay-21.01" location="cell-11" area-ref="flegg-page-frame"
                     valign="right"/>
        <layout-leaf xref="lay-21.02" location="cell-12" area-ref="flegg-page-frame"/>
      </layout-chunk>
      <layout-chunk id="title" location="cell-11" area-ref="flegg-page-frame">
        <layout-leaf xref="lay-21.03" location="col-1" area-ref="title-cols"
                     halign="bottom"/>
        <layout-leaf xref="lay-21.04" location="col-2" area-ref="title-cols"
                     halign="bottom"/>
      </layout-chunk>
      <layout-leaf xref="line-21.1" location="delimiter" delimiter-after="cell-11"
                   delimiter-before="cell-21" area-ref="flegg-page-frame"/>
      <layout-chunk id="flegg-body" location="multi" area-ref="flegg-page-frame">
        <layout-leaf xref="flegg-photo" location="row-2" area-ref="flegg-page-frame"/>
        <layout-leaf xref="flegg-text" location="cell-31" area-ref="flegg-page-frame"/>
        <layout-chunk id="flegg-table" location="cell-41" area-ref="flegg-page-frame">
          <layout-leaf xref="lay-21.12" location="cell-11" area-ref="table-frame"
                       valign="right"/>
          <layout-leaf xref="lay-21.13" location="cell-12" area-ref="table-frame"/>
          <layout-leaf xref="lay-21.14" location="cell-21" area-ref="table-frame"
                       valign="right"/>
          <layout-leaf xref="lay-21.15" location="cell-22" area-ref="table-frame"/>
          <layout-leaf xref="lay-21.16" location="cell-31" area-ref="table-frame"
                       valign="right"/>
          <layout-leaf xref="lay-21.17" location="cell-32" area-ref="table-frame"/>
          <layout-leaf xref="lay-21.18" location="cell-41" area-ref="table-frame"
                       valign="right"/>
          <layout-leaf xref="lay-21.19" location="cell-42" area-ref="table-frame"/>
          <layout-leaf xref="lay-21.20" location="cell-51" area-ref="table-frame"
                       valign="right"/>
          <layout-leaf xref="lay-21.21" location="cell-52" area-ref="table-frame"/>
        </layout-chunk>
      </layout-chunk>
      <layout-leaf xref="line-21.2" location="cell-51" area-ref="flegg-page-frame"/>
      <layout-leaf xref="flegg-page-no" location="cell-52" area-ref="flegg-page-frame"/>
    </layout-root>
  </layout-structure>
</gemLayout>
```

Appendix 3: XML file of the RST structure

```
<?xml version="1.0"?>
<!DOCTYPE gemRst SYSTEM ".../GeM/corpus/dtds/gem-rst.dtd">
<gemRst>
<segmentation>
   <mini-segment id="s-21.03" xref="u-21.03">GANNET</mini-segment>
   <mini-segment id="s-21.04" xref="u-21.04">Sula bassana</mini-segment>
  <segment id="s-21.06" xref="u-21.06"/>
  <segment id="s-21.07" xref="u-21.07">Huge (90cm) unmistakable seabird.</segment>
  <segment id="s-21.08" xref="u-21.08">Watch for white, cigar-shaped body and long
          straight, slender, black-tipped wings.</segment>
  <segment id="s-21.09" xref="u-21.09">In summer, yellow head of adult inconspicuous.
  </segment>
  <segment id="s-21.10" xref="u-21.10">Plunges spectacularly for fish.</segment>
  <segment id="s-21.11" xref="u-21.11">Sexes similar. </segment>
  <mini-segment id="s-21.12" xref="u-21.12">Juvenile </mini-segment>
  <mini-segment id="s-21.13" xref="u-21.13">Grey-brown, flecked becoming whiter, adult
               plumage after three years.</mini-segment>
  <segment id="s-21.15" xref="u-21.15">Breeds colonially on cliffs on N and W coasts,
          dispersing to winter at sea. </segment>
  <mini-segment id="s-21.16" xref="u-21.16">Nest</mini-segment>
  <mini-segment id="s-21.17" xref="u-21.17">Mound of seaweed on bare rocky ledge.
  </mini-segment>
  <mini-segment id="s-21.18" xref="u-21.18">Voice</mini-segment>
  <mini-segment id="s-21.19" xref="u-21.19">Harsh honks and grating calls at colony.
  </mini-segment>
  <segment id="s-21.20" xref="u-21.20">General</segment>
  <segment id="s-21.21" xref="u-21.21">Widespread, but breeding colonies few though
          sometimes enormous.</segment>
</segmentation>
<rst-structure root=‘‘flegg-gannet’’>
  <span id="span-21.31" nucleus="s-21.07" satellites="s-21.08 s-21.09 s-21.10"
        relation="justify"/>
  <span id="span-21.32" nucleus="span-21.31" satellites="s-21.11" relation="background"/>
  <multi-span id="flegg-recognition-21" nuclei="s-21.06 span-21.32" relation="joint"/>
  <multi-span id="flegg-habitat-21" nuclei="s-21.24 s-21.15 s-21.25 s-21.26 span-21.34"
             relation="joint"/>
  <span id="span-21.34" nucleus="s-21.21" satellites="s-21.20" relation="interpretation"/>
  <span id="span-21.1" nucleus="flegg-recognition-21" satellites="flegg-habitat-21"
        relation="elaboration"/>
  <span id="flegg-gannet" nucleus="span-21.1" satellites="s-21.22" relation="background"/>
</rst-structure>
<mini-structure>
  <mini-span id="s-21.22" attribuend="s-21.03" attribute="s-21.04"
             relation="identification">GANNET  Sula bassana</mini-span>
  <mini-span id="s-21.24" attribuend="s-21.12" attribute="s-21.13"
             relation="property-ascription">Juvenile Grey-brown, flecked becoming whiter,
             adult plumage after three years.</mini-span>
  <mini-span id="s-21.25" attribute="s-21.16" attribuend="s-21.17"
             relation="class-ascription">Nest Mound of seaweed on bare rocky ledge.
  </mini-span>
  <mini-span id="s-21.26" attribute="s-21.18" attribuend="s-21.19"
```

```
            relation="class-ascription">Voice Harsh honks and grating calls at colony.
  </mini-span>
</mini-structure>
</gemRst>
```

Appendix 4: XML file of the navigation structure

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE gemNav SYSTEM "../..//dtds/gem-nav.dtd">
<!-- finished: 11.09.2002 -->
<!-- corrections: 06.02.2003 -->
<gemNav>
<entry id="e-21.01" xref="u-21.01" rst-span="external"/>
<entry id="e-21.02" xref="u-21.02" rst-span="external"/>
<entry id="e-21.03" xref="u-21.03 u-21.04" rst-span="flegg-gannet"/>
<entry id="e-21.14" xref="u-21.14" rst-span="s-21.15"/>
<index id="i-21" name="21" layout-chunk="flegg-page-21" xref="u-21.23">21</index>
</gemNav>
```