GeM Project Report
University of Stirling
University of Bremen

http://www.gem.stir.ac.uk
fax.: fax?
e-mail: henschel@uni-bremen.de

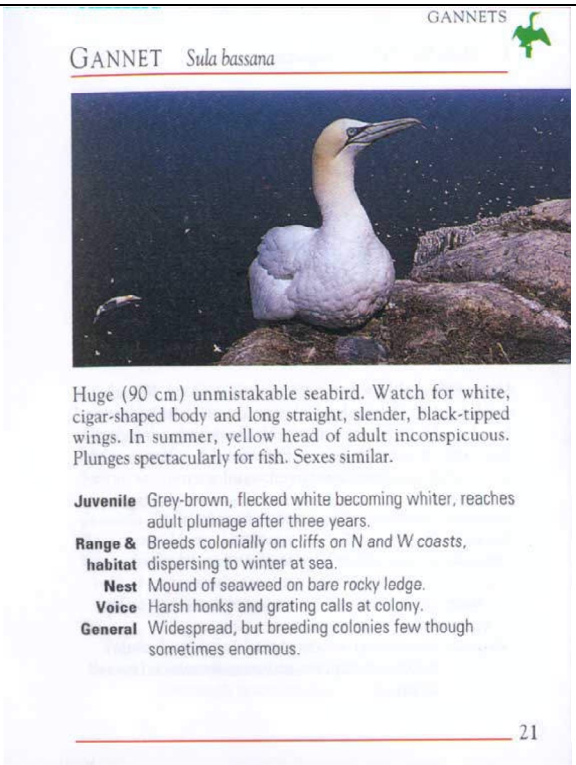
GeM Annotation: One complete example – the Flegg page

Renate Henschel

December 17, 2001

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 and the tutorial. 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 appendixes; 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-21.1	GANNETS
u-21.2	(gannet icon)
u-21.3	GANNET
u-21.4	Sula Bassana
u-21.5	(horizontal red line)
u-21.6	(photo)
u-21.7	Huge (90cm) unmistakable seabird.
u-21.8	Watch for white, cigar-shaped body and long straight, slender, black-tipped wings.
u-21.9	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 and tutorial: 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-21.7 until u-21.11), form together a single unit as far as the layout is concerned. The remaining 18 base units are function as 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. When an element is unique for its type on the page, or for other mnemonic reasons, we may as required deviate from the strict numbering system and give a name. 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 labelling, each layout unit has a unique identifier (e.g., "lay-21.3") and a reference to the base unit or base units to which it corresponds (e.g., for "lay-21.3" the base unit "u-21.3"). This is represented in the XML annotation with lines such as:

```
<layout-unit id="lay-21.3" href="u-21.3"/>
```

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-21.2), 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. The photo is of type="photo". For photos and illustrations, we choose additionally between the color values black-white or color. In this case it would be color="color". For the red lines, we also mark 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. An example annotation for the photo is therefore:

```
<graphics href="flegg-photo" type="photo" color="color" height="5.6cm"/>
```

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. In contrast, 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. An example annotation for the typographical features is:

```
<text href="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"/>
```

The other text elements all differ from each other, and so receive their own text element annotation; such as, for example, the following:

```
<text href="lay-21.3" font-family="serif" font-size="14"
      font-style="normal"
      font-weight="normal" case="caps"
      justification="no" color="black''>
  GANNET
</text>
```

It is not quite clear in this page, whether the page number shares its typographical layout with flegg-text. We have chosen 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 humber).

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!

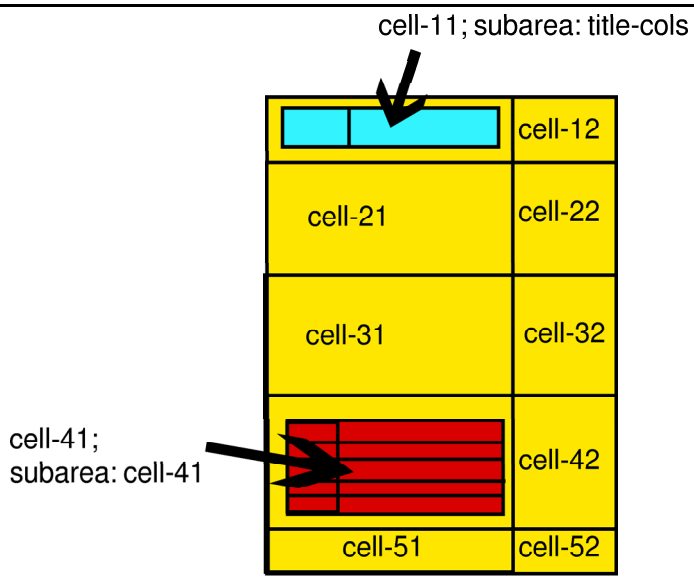


Figure 2: Area model for the flegg page showing subareas

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 or then indicated by percentages as suggested above. For example, for the case of the titles, 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 table. The table itself has 10 layout units, the actual text in the einzelnen table cells. The two red lines support 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 possibility would be to perceive them as part of the page body. The resulting layout structure is then as shown in Figure 3.

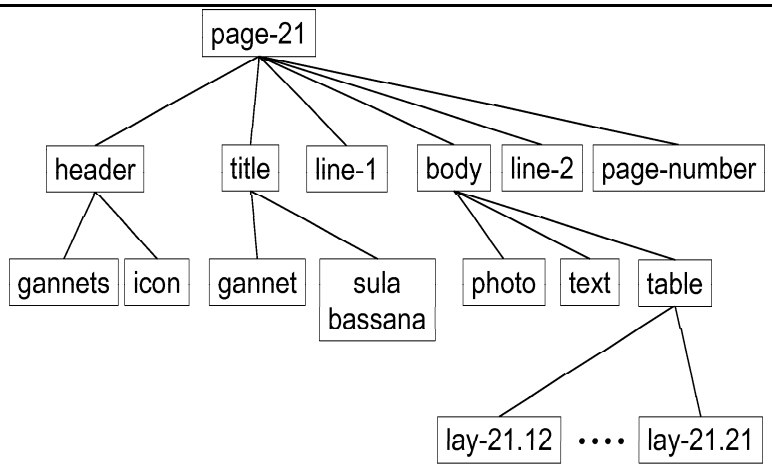


Figure 3: Layout structure for the flegg page

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 with 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 system, 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 model "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".

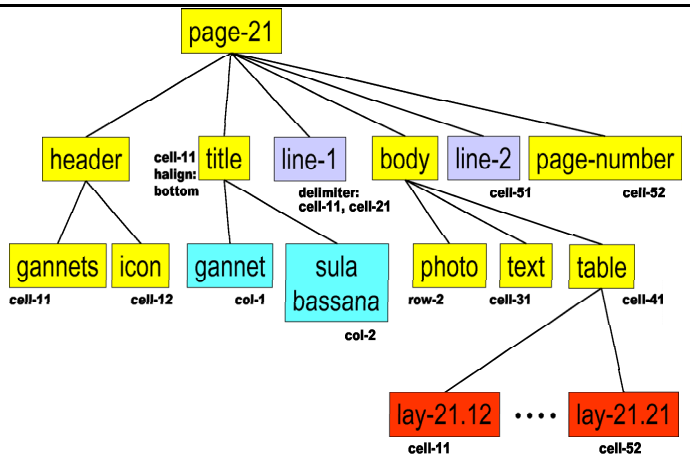


Figure 4: Layout structure for the flegg page with locations

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.

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

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-21.1">GANNETS</unit>
  <unit id="u-21.2"/>                        <!-- icon -->
  <unit id="u-21.3">GANNET</unit>
  <unit id="u-21.4">Sula Bassana</unit>
  <unit id="u-21.5">----------------------------------------</unit>
  <unit id="u-21.6"/>                        <!-- photo -->
  <unit id="u-21.7">Huge (90cm) unmistakable seabird.</unit>
  <unit id="u-21.8">Watch for white, cigar-shaped body and long straight,
```

```
                    slender, black-tipped wings. </unit>
  <unit id="u-21.9">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-21.1" href="u-21.1"/>
    <layout-unit id="lay-21.2" href="u-21.2"/>
    <layout-unit id="lay-21.3" href="u-21.3"/>
    <layout-unit id="lay-21.4" href="u-21.4"/>
    <layout-unit id="line-21.1" href="u-21.5"/>
    <layout-unit id="flegg-photo" href="u-21.6"/>
    <layout-unit id="flegg-text" href="u-21.7 u-21.8 u-21.9 u-21.10 u-21.11"/>
    <layout-unit id="lay-21.12" href="u-21.12"/>
    <layout-unit id="lay-21.13" href="u-21.13"/>
    <layout-unit id="lay-21.14" href="u-21.14"/>
    <layout-unit id="lay-21.15" href="u-21.15"/>
    <layout-unit id="lay-21.16" href="u-21.16"/>
    <layout-unit id="lay-21.17" href="u-21.17"/>
    <layout-unit id="lay-21.18" href="u-21.18"/>
    <layout-unit id="lay-21.19" href="u-21.19"/>
    <layout-unit id="lay-21.20" href="u-21.20"/>
    <layout-unit id="lay-21.21" href="u-21.21"/>
    <layout-unit id="line-21.2" href="u-21.22"/>
    <layout-unit id="flegg-page-no" href="u-21.23"/>
  </segmentation>
  <realization>
    <!-- textual elements with similar typographical features -->
    <text href="lay-21.1" font-family="serif" font-size="8" font-style="normal
          font-weight="normal" case="caps" justification="no" color="black">GANNETS</text>
```

```
  <text href="lay-21.3" font-family="serif" font-size="14" font-style="normal"
        font-weight="normal" case="caps" justification="no" color="black''>GANNET</text>
  <text href="lay-21.4" font-family="serif" font-size="12" font-style="normal"
        font-weight="normal" case="caps" justification="no" color="black">
           Sula bassana</text>
  <text href="flegg-text" font-family="serif" font-size="11" font-style="normal"
        font-weight="normal" case="mixed" justification="justified" color="black"/>
  <text href="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 href="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 href="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 href="lay-21.2" type="two-d-element" two-d-element-type="icon" color="green"/>
  <graphics href="flegg-photo" type="photo" color="color" height="5.6cm"/>
  <graphics href="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 href="lay-21.1" location="cell-11" area-ref="flegg-page-frame"
                   valign="right"/>
      <layout-leaf href="lay-21.2" location="cell-12" area-ref="flegg-page-frame"/>
    </layout-chunk>
    <layout-chunk id="title" location="cell-11" area-ref="flegg-page-frame">
      <layout-leaf href="lay-21.3" location="col-1" area-ref="title-cols"
                   halign="bottom"/>
      <layout-leaf href="lay-21.4" location="col-2" area-ref="title-cols"
                   halign="bottom"/>
    </layout-chunk>
    <layout-leaf href="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 href="flegg-photo" location="row-2" area-ref="flegg-page-frame"/>
      <layout-leaf href="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 href="lay-21.12" location="cell-11" area-ref="table-frame"
                     valign="right"/>
        <layout-leaf href="lay-21.13" location="cell-12" area-ref="table-frame"/>
        <layout-leaf href="lay-21.14" location="cell-21" area-ref="table-frame"
                     valign="right"/>
```

```
          <layout-leaf href="lay-21.15" location="cell-22" area-ref="table-frame"/>
          <layout-leaf href="lay-21.16" location="cell-31" area-ref="table-frame"
                       valign="right"/>
          <layout-leaf href="lay-21.17" location="cell-32" area-ref="table-frame"/>
          <layout-leaf href="lay-21.18" location="cell-41" area-ref="table-frame"
                       valign="right"/>
          <layout-leaf href="lay-21.19" location="cell-42" area-ref="table-frame"/>
          <layout-leaf href="lay-21.20" location="cell-51" area-ref="table-frame"
                       valign="right"/>
          <layout-leaf href="lay-21.21" location="cell-52" area-ref="table-frame"/>
        </layout-chunk>
      </layout-chunk>
      <layout-leaf href="line-21.2" location="cell-51" area-ref="flegg-page-frame"/>
      <layout-leaf href="flegg-page-no" location="cell-52" area-ref="flegg-page-frame"/>
    </layout-root>
  </layout-structure>
</gemLayout>
```