Quy định đồ án

1. Mục tiêu

Mục tiêu của đồ án là xây dựng một **Digital Talking Book (DTB)** theo chuẩn **DAISY v3**, với nội dung văn bản trích xuất từ PDF và âm thanh thu thập từ **Thư viện Sách Nói Hướng Dương**.

Sinh viên phải đảm bảo:

- Cố gắng thực hiện các quy trình **xử lý tự động ở mức cao nhất có thể**.
- Đồng bộ chính xác giữa văn bản và audio.
- Cấu trúc tệp và thư mục tuân thủ chuẩn DAISY v3.
- Metadata đầy đủ và chính xác.
- Báo cáo quá trình thực hiện, sai lệch và giải pháp xử lý.

---

2. Tài liệu tham khảo bắt buộc

Sinh viên phải đọc và tuân thủ các tài liệu chuẩn:

1. **DAISY 3 Standard (Z39.86-2005 R2012)**

   https://daisy.org/activities/standards/daisy/daisy-3/z39-86-2005-r2012-specifications-for-the-digital-talking-book/

2. **DAISY 3 Structure Guidelines (SG-DAISY3)**

   https://dl.daisy.org/standards/z3986/SG-DAISY3.zip

---

3. Yêu cầu thực hiện

3.1. Thu thập dữ liệu

- Tải audio từ **thuviensachnoihuongduong.com** theo sách được phân công.
- PDF cung cấp sẵn để trích xuất văn bản.

3.2. Chuẩn hóa văn bản

- Loại bỏ số trang, header/footer, watermark, liên kết dư thừa và các ký tự không liên quan.

- Chuẩn hóa Unicode và khoảng trắng.
- Chia văn bản thành các **đơn vị đồng bộ** (câu hoặc đoạn) để thống nhất toàn bộ.

3.3. Đồng bộ văn bản – âm thanh

- Phân đoạn audio bằng **SMIL** với `clipBegin` và `clipEnd` chính xác.
- Xử lý sai lệch: audio thiếu, thừa, sai nội dung.

3.4. Xây dựng cấu trúc DAISY v3

Sản phẩm phải có:

- **main.xml** nội dung văn bản
- **.smil** đồng bộ
- **main.opf** chứa metadata, manifest, spine và liên kết tài nguyên
- **.ncx** điều hướng (nếu có)
- **.mp3** audio

---

4. Metadata bắt buộc

Trường	Mô tả	Bắt buộc
title	Tên sách	Có
creator	Tác giả	Không
subject	Tên thể loại sách	Có
description	Mô tả sơ lược sách	Không
publisher	Nhà phát hành	Có
date	Ngày phát hành, định dạng: yyyy, yyyy-mm, yyyy-mm-dd	Có
source	Mã ISBN của sách	Có
language	Ngôn ngữ	Có
note*	Ghi chú thêm	Không
collector*	Tên người đóng góp sách	Không
sourceURL*	URL gốc sách	Không

Trường	Mô tả	Bắt buộc
thumb*	Ảnh bìa sách	Có

main.xml

```
<head>
      <meta content="Cánh đồng bất tận" name="dc:Title"></meta>
      <meta content="Nguyễn Ngọc Tư" name="dc:Creator"></meta>
      <meta content="2006" name="dc:Date"></meta>
      <meta content="Cánh đồng bất tận là tên một tập truyện ngắn phát hà
nh năm 2005 của Nguyễn Ngọc Tư, đồng thời cũng là tên một truyện trong
tập truyện ngắn đó được đăng báo lần đầu cùng năm. Tính đến nay, tập tru
yện đã được phát hành dưới dạng sách in và sách nói." name="dc:Descripti
on"></meta>
      <meta content="vi" name="dc:Language"></meta>
      <meta name="dc:Subject" content="Văn học &#x26; Tiểu thuyết"></
meta>
      <meta name="dc:Publisher" content="NXB Trẻ"></meta>
      <meta name="dc:Source" content="9786041266001"></meta>
      <meta name="thumb" content="https://encrypted-tbn0.gstatic.com/i
mages?q=tbn:ANd9GcTjMjLWbUH4ndoOblh57PCSQTovy9mcExFpZg&s">
</meta>
   </head>
```

main.opf

```
<metadata>
   <dc-metadata>
      <dc:Title
         xmlns:dc="http://purl.org/dc/elements/1.1/">Cánh đồng bất tận
      </dc:Title>
      <dc:Creator
         xmlns:dc="http://purl.org/dc/elements/1.1/">Nguyễn Ngọc Tư
      </dc:Creator>
      <dc:Date
         xmlns:dc="http://purl.org/dc/elements/1.1/">2006
      </dc:Date>
      <dc:Description
         xmlns:dc="http://purl.org/dc/elements/1.1/">Cánh đồng bất tận là tê
```

```
n một tập truyện ngắn phát hành năm 2005 của Nguyễn Ngọc Tư, đồng thờ
i cũng là tên một truyện trong tập truyện ngắn đó được đăng báo lần đầu cù
ng năm. Tính đến nay, tập truyện đã được phát hành dưới dạng sách in và s
ách nói.
      </dc:Description>
      <dc:Language
        xmlns:dc="http://purl.org/dc/elements/1.1/">vi
      </dc:Language>
      <dc:Subject
        xmlns:dc="http://purl.org/dc/elements/1.1/">Văn học &#x26; Tiểu th
uyết
      </dc:Subject>
      <dc:Publisher
        xmlns:dc="http://purl.org/dc/elements/1.1/">NXB Trẻ
      </dc:Publisher>
      <dc:Source
        xmlns:dc="http://purl.org/dc/elements/1.1/">9786041266001
      </dc:Source>
    </dc-metadata>
    <x-metadata>
      <meta name="thumb" content="https://encrypted-tbn0.gstatic.com/i
mages?q=tbn:ANd9GcTjMjLWbUH4ndoOblh57PCSQTovy9mcExFpZg&s"/>
    </x-metadata>
</metadata>
```

---

5. Cấu trúc thư mục nộp

```
MSHV1_MSHV2_MSHV3/
├── Tên_sách_1-Chương_1/
│   ├── Tên_sách.zip            ← Gồm .xml, .smil, .mp3, .opf, .ncx nếu có
│   └── Tên_sách_sha256sums.txt   ← Hash SHA-256
├── Tên_sách_1-Chương_2/
│   ├── Tên_sách.zip
│   └── Tên_sách_sha256sums.txt
└── Tên_sách_2/
```

```
    ├── Tên_sách.zip
    └── Tên_sách_sha256sums.txt
```

> Đối với các cuốn sách có kích lớn như, có thể chia thành các phần nhỏ hơn như chương, hồi.