

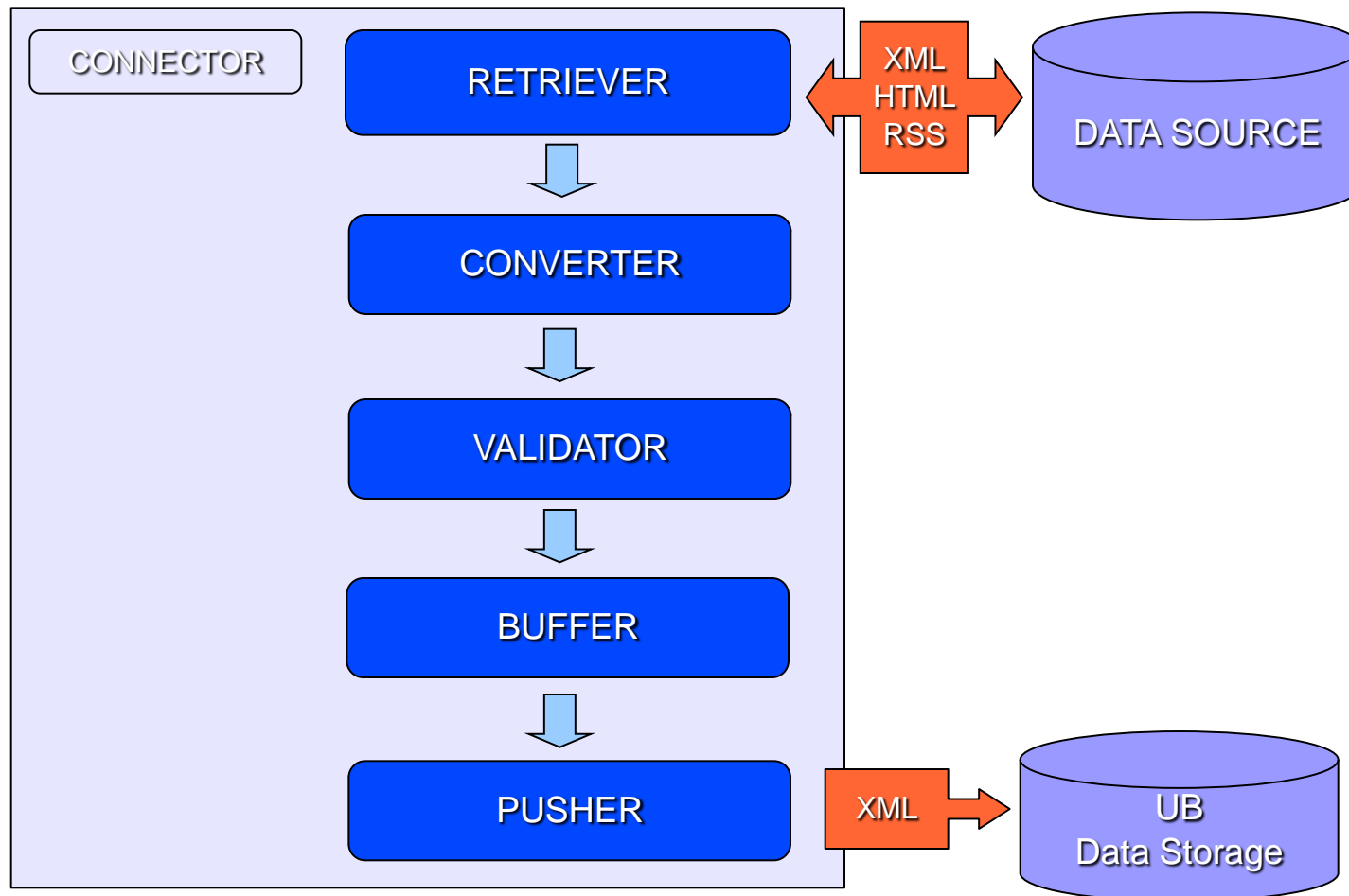
Pobieranie danych NetSprint NASP

Sposoby pobierania danych w systemie NASP:

- **Connector**
 - dane wystawione przez Klienta w postaci plików XML
 - dane pobierane bezpośrednio z baz danych
 - dokumenty z serwerów FTP
 - dokumenty biurowe z lokalizacji dyskowych
- **WebSpider (Crawler)**
 - treści bezpośrednio ze stron WWW

Connector

- **Źródła danych** (system plików, baza danych, serwer FTP)
- **Format danych wejściowych** (XML, HTML, RSS, MSOffice, PDF)
- **Moduły Connectora**
 - Retriever
 - Converter
 - Validator
 - Buffer
 - Pusher



Funkcjonalności Connectora:

- Obsługa protokołu HTTP, FTP
- Autoryzacja: hasło, login + hasło, inne
- Pobieranie dokumentów z określonych przedziałów czasowych
- Konfigurowalna strategia odświeżania dokumentów
- Konfigurowalna strategia usuwania dokumentów
- Możliwość zdefiniowania dowolnej liczby pól opisujących dokument
- Równoczesne zasilanie wielu magazynów (UB)
- Równoczesny zapis do magazynu oraz pliku backupowego

Przykładowe parametry:

- Source
- DateFrom
- DateTo
- Pass
- Login

Przykładowe zapytanie:

<http://adresIP/DocumentsList?Source=1&DateFrom=20100104&DateTo=20100108&Pass=2Z>

Przykładowa struktura dokumentów przeznaczonych do wstawienia do systemu:

```
<?xml version="1.0" encoding="UTF-8"?>  
<doc>  
  <id> id dokumentu </id>  
  <source> źródło pochodzenia dokumentu</source>  
  <title> tytuł dokumentu </title>  
  <body> treść dokumentu </body>  
  <category> kategoria dokumentu </category>  
  <pubDate> data publikacji dokumentu</pubDate>  
</doc>
```

Przykładowa struktura dokumentów przeznaczonych do usunięcia z systemu:

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<documents>
```

```
<doc>
```

```
    <id> id dokumentu </id>
```

```
</doc>
```

```
<doc>
```

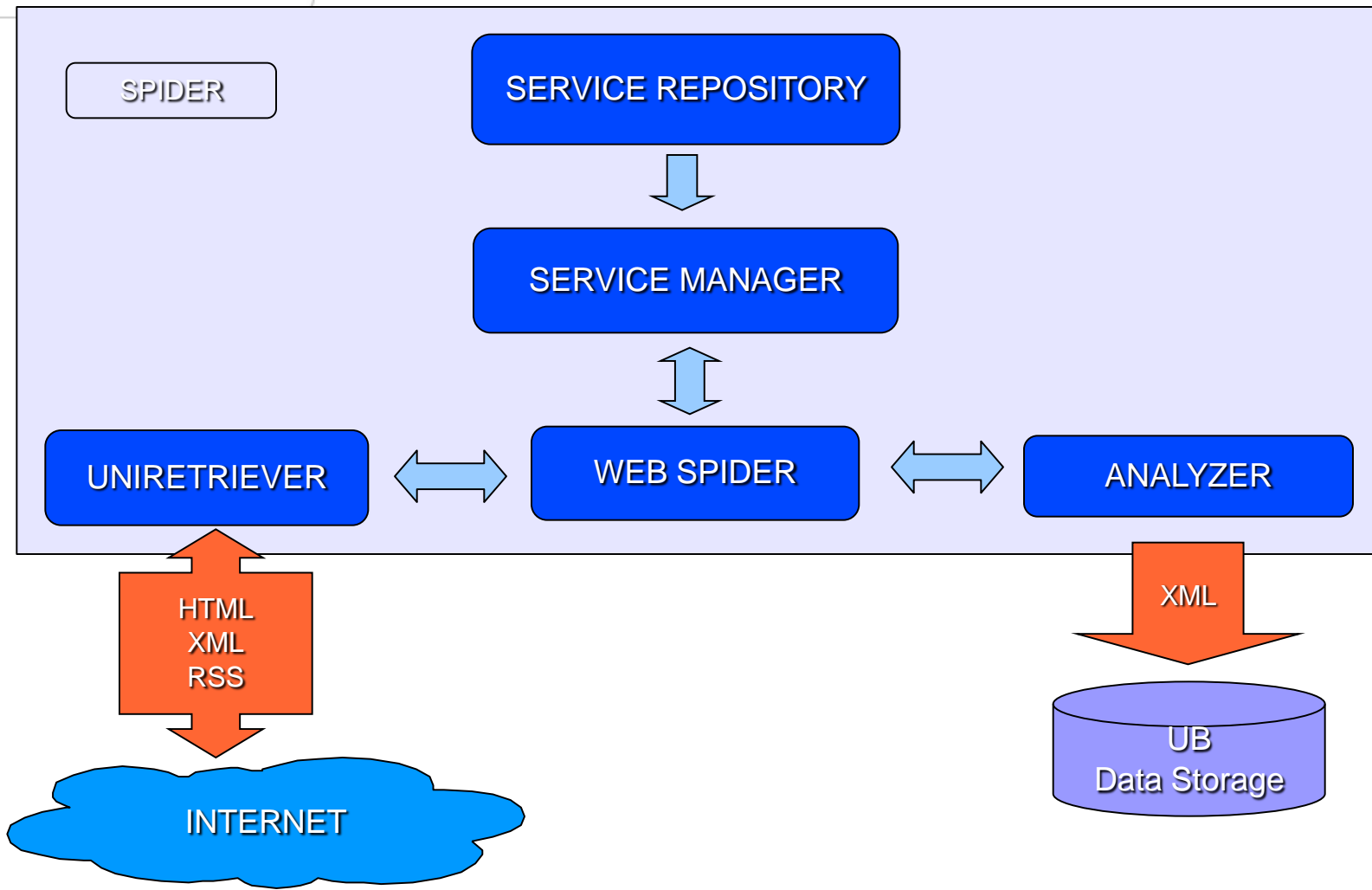
```
    <id> id dokumentu </id>
```

```
</doc>
```

```
</documents>
```

Spider

- **Źródła danych** (serwisy internetowe: HTTP, RSS)
- **Format danych wejściowych** (XML, HTML, RSS)
- **Moduły Spidera**
 - Service Repository
 - Service Manager
 - Web Spider
 - Uniretriever
 - Analityzer



Funkcjonalności Spidera:

- Obsługa protokołu HTTP
- Autoryzacja: stałe cookie, login + hasło, inne
- Filtr zbieranych linków
- Limit głębokości zbierania
- Konfigurowalna strategia odświeżania dokumentów
- Możliwość zdefiniowania dowolnej liczby pól opisujących dokument
- Równoczesne zasilanie wielu magazynów (UB)
- Równoczesny zapis do magazynu oraz pliku backupowego

Informacje potrzebne do konfiguracji Spider-a dla serwisu:

- URL startowy
- Linki do następnych stron
- Odstęp czasowy
- Struktura zbieranych dokumentów
- Zakres zbieranych informacji

Informacje pamiętane dla serwisu:

- Historia ostatnio odwiedzonych linków
- Data ostatniego odwiedzenia

Pierwsze pobranie danych:

- Pobranie wszystkich typów dokumentów ze wszystkich zdefiniowanych źródeł danych
- Struktura dokumentów powinna być zgodna ze specyfikacją zawartą w dokumentacji technicznej
- Pliki mogą zostać udostępnione poprzez:
 - HTTP
 - FTP
 - System plików

