2005－SE－147（6）
2005／3／17

文脈に基づいたソースプログラムとドキュメント間の識別子対応付け手法

後藤 英斗　　大久保 弘崇　　粕谷 英人　　山本 晋一郎

愛知県立大学 情報科学部 情報システム学科

要 旨

ソースプログラムとXMLで記述されているドキュメントに含まれる識別子を自動的に対応付ける手法を提案する．ドキュメント中の識別子の出現は文脈から定義か参照かを判断する．対応付けの応用として，ソフトウェアの理解支援のためにソースプログラムとドキュメントの相互参照を提供するツールと，両者の整合性を検査するツールを実装し，その有用性を確認した．

A Method for Relating Identifiers between Source Programs and Documents using their Contexts.

Hideto GOTO, Hirotaka OHKUBO, Hideto KASUYA, and Shinichiro YAMAMOTO

Department of Information Systems, Faculty of Information Science,Aichi Prefectual University

Abstract

In this paper, we propose a method for relating identifiers appear in source programs and documents such as specifications and manuals. Our method distinguishes between definitions and references of identifiers in documents using their contexts. Generated correspondences can be applicable to CASE tools such as cross referencer across source programs and documents or consistency checker between them.

1　はじめに

1.1　背景

ソフトウェアの開発や再利用にあたり，多くの場合，作業者はソースプログラムを理解する必要があるが，モジュールの呼び出し関係が複雑であったり，各種識別子の役割がわからないなどさまざまな理由により，その作業は困難である．そこで，ソースプログラムを解析してプログラマの理解を支援するためのツールが提案されている．例えばCASEツールプラットフォームSapid [1, 2] では，ソフトウェアモデルに基づいた解析を行いソフトウェアデータベースSDBを生成する．このSDBをアクセスルーチンARを用いて利用することで解析結果に基づいたソフトウェアの理解を支援する機能を提供する．また，JavadocやDoxygenのようなソースプログラム中にドキュメントを簡潔に記述するための枠組も提案されている．

しかし，仕様書やライブラリのマニュアルなどは別ファイルに保存されるのが一般的である．また，開発者間でやりとりされるメールなどのように，ソースプログラムに埋め込むことは難しい．このように，同一のソフトウェアを

構成するソースプログラムと関連ドキュメントが別々のファイルにそれぞれ記述されているため，相互に連携して利用することは現状では困難である．また，ソースプログラムを対象としたツールに比較して，ドキュメントを対象として解析や参照を支援するツールも少ない．

1.2 目的

我々は,「ソフトウェア＝ソースプログラム＋ドキュメント」という観点で両者の密接な連携を支援するための研究を行っている．ソースプログラムとドキュメントは同じ１つのソフトウェアの構成要素であるが，別々のファイルに存在しているためにそれぞれ別の用途のためのものと考えられがちである．そこで，別々のファイルに記述されるソースプログラムとドキュメントを対応付けることによって連携して活用する方法を提案する．本研究の対象はソースプログラムとドキュメント中に現れる識別子の対応づけに留まっているが，これは将来の研究課題である，ドキュメント中のある段落とそれを実現しているソースプログラム部分との対応付けに代表される，より抽象度が高く有用な対応づけへの第一歩である．

例えば，ソースプログラムの閲覧時に現在注目している関数の仕様が知りたい場合，改めてドキュメントのファイルを開いて，検索機能や索引を用いて該当箇所を探さなくてはならない．その作業は決して難しいことではないが回数が増えると煩雑になり，作業効率にも影響を及ぼす．この煩雑な参照作業は予め両者間を対応付けておくことで解決できる．対応付けによってリンクを作成しておけば，検索機能などを用いることなく参照が行える．

1.3 概要

本稿ではソースプログラムとドキュメントのそれぞれに含まれる識別子に注目し，これらを対応付けて利用する手法を次の 2 に分けて提案する．

- ソースプログラムとドキュメントの識別子の対応付け
- 対応付けを応用したツール

また，ドキュメントに含まれる識別子に関して，文脈の考慮を提案する．ソースプログラム中の識別子の出現を宣言と参照に分けられるように，これを「定義」と「参照」に区別して扱う方法である．

対応付けの目的は利用者により様々であると考えられるので，対応付けとその応用を別々の手法として提案する．利用者はまずソースプログラムとドキュメントの対応付けを行い，その結果を目的に合った応用ツールで利用する．

2 章ではソースプログラムとドキュメントの対応に関する研究を紹介する．3 章で対応付けの手法について説明をし，4 章でその対応付けを応用するツールを 2 つ紹介する．

2 ソースプログラムとドキュメントの対応に関する研究

2.1 ADIOS

ADIOS [5, 6] はソフトウェアに関係する多種のドキュメントをアスペクト指向を用いて整理する手法を提案している．ソースコード中のキーワードからドキュメントへの参照を「関連」と呼び，その関連をソースコード中に一定の形式で埋め込む手法を用いている．埋め込まれた関連はこの手法を実装したツールである ADIOS によって収集され，これをアスペクト指向により横断的に整理する．関連には種類やキーワードが設定してあり，これらをもとに整理を行い同類の関連をまとめることで，関連先を推移的に辿ることができる．

ソースプログラム中には関連のみを埋め込む手法であるため，後述の Javadoc や Doxygen とは異なりソースプログラムの可読性に影響を与える可能性は少ない．関連を手動で埋め込む必要はあるが，後から収集して整理することができるので利用性は高い．

2.2 机上デバッグ支援ツール Prio

Prio [7] に関する研究では仕様書に記述されている情報とプログラムに記述されている情

報の間の不整合を検査することによってソフトウェアの不具合を発見することを目的としている．このツールを用いた机上チェックを行うことにより，仕様書とプログラムの不整合によるミスを効率的に発見できる．また，後工程で評価作業時間を短縮することもできる．

この研究はプログラムの不具合を仕様書の不整合から見つけようという試みである．対応関係を示した XML 文書を基にソースプログラムと仕様書の間での矛盾を検出する．ソースプログラムのみでは見つけにくい不具合でも，仕様書と食い違っていることによって発見できれば成果物の品質を向上させられる．

2.3 文芸的プログラミング

文芸的プログラミング [8] とは，Donald Knuth が提唱したプログラムの作成とその文書化を同時に行う手法である．「コンピュータプログラムは計算機が実行可能でなくてはならないが，それは主な目的ではない．有用なコンピュータプログラムは人間が読むことができなくてはならない」という考えに基づいている．

プログラムとその文書をひとつのファイルに混在させながら追加したり修正したりして同時に開発していくことにより、完全に整合性のとれた文書とプログラムの開発を目指している．プログラムをまず作ってからその解説文書を作成する (またはその逆) という一般的な手法では、後で修正などを加える場合などは注意しないと両者の整合性がとれなくなってしまうことがよくあるが、文芸的プログラミングの手法では両者がひとつのファイルになっているのでそのようなことが起こりにくい．

2.4 Javadoc, Doxygen, Perldoc

ソースプログラムのコメント中に規定の形式でドキュメントを埋め込む方法を持つプログラミング言語がある．Java では Javadoc，C++では Doxygen，Perl では perldoc がこのドキュメントを抽出するツールである．

Javadoc クラス，メソッド，フィールドの直前にコメントとしてそれぞれの説明を記述する．例えば “`@param` 引数名 コメント” という記法でメソッドの引数の説明を記述する．`javadoc` コマンドでこれらの記述を HTML ファイル群にして取り出せる．記述の中に HTML 断片を埋め込むこともできる．

Doxygen 記法は Javadoc と似ている．`doxygen` コマンドによって HTML ファイルをはじめ，TEX や RTF のファイルも生成できる．

Perldoc perl プログラムのファイルには POD という形式で説明を記述できる．この説明記述は `perldoc` コマンドによって取り出し，`pod` コマンド群によって man 形式や HTML 形式などに変換して参照できる．

2.5 関連研究の考察

ADIOS や Prio は別個にファイルに記述されたソースプログラムとドキュメントを連携して利用できる．このツールを用いることでソースプログラムとドキュメントの対応関係を集約してソフトウェア開発に利用できる．しかし，対応を人手で指定しなくてはならない点が短所である．ソースプログラムの作成とドキュメントの作成のほかに，ソースプログラムとドキュメントの対応づけをするという作業が必要となる．

文芸的プログラミングは，実際のコードは自然言語で記述された文書の部分が圧倒的に多く，それに対して実行コードに変換される部分は僅かしかない．Knuth も著書 [8] のなかで「プログラムを文学と捉える」,「プログラムは人間に読めなくてはならない」と述べているので，そのようになるのは当然である．文芸的にプログラムするということは先に自然言語の文書を作成する必要がある．しかしプログラミングにおいてはプログラム作成中に仕様や方針が変わるということは多いので，このような場合に文書も変更する必要があり不便である．

Javadoc などのソースプログラムのコメントにドキュメントを埋め込む手法ではソースプログラム中の説明を行う要素の近くに説明記述が存在するため，同時に修正を施すことができ，ソースプログラムとドキュメントの間の整合性

を保つことに長けている．しかし，説明記述が増えるとソースプログラムのファイルが巨大になり，またファイルをプログラムのコードよりも文章のほうが多くを占めるようになるのでソースプログラムの可読性に影響する．

ADIOS，Prio は対応関係を作成者が手作業で設定させる手間が短所である．この作業を自動で行うことができれば利用しやすくなる．一方，文芸的プログラミングやソースプログラムにドキュメントを埋め込む手法はソースプログラムのファイルが巨大化，複雑化することが短所である．この点も ADIOS や Prio の短所と同様にソースプログラムとドキュメントの間の対応箇所を自動で抽出できるようにすることで解決できる．１つのファイルの中にソースプログラムとドキュメントを混在させなくても，抽出した対応関係を用いて相互に参照できればソースプログラムとその説明記述が近くにあるのと同様の効果が得られる．

3　識別子の対応付け

本稿では，ソースプログラムとドキュメントのそれぞれで種類と名前が同じ識別子の出現を見つけて組にする作業を「対応付け」と呼び，この組を「対応」と呼ぶ．ソースプログラム，ドキュメントそれぞれで同じ識別子は複数回出現する．従って，ソースプログラム中の１つの識別子の出現に対してドキュメント中の複数の出現が対応付けられる．また，逆にドキュメント中の１つの識別子の出現に対してもソースプログラム中の複数の出現が対応付けられる．

入力ファイルは識別子の抽出のためにソースプログラム，ドキュメントともに XML でマークアップしたファイルを対象とする．ソースプログラムに対する意味解析は予め行われているものとして，その結果にしたがって SPIE よってマークアップされたものを用いる．ドキュメントは DocBook と呼ばれる形式で記述されたものを対象とする．

3.1　ソースプログラムのマークアップ

ソースプログラムは SPIE [4] によってマークアップされたものを扱う．SPIE は CASE ツールプラットフォーム Sapid [1, 2] に含まれるツールであり，Sapid によるソースプログラムの解析に基づいてソースプログラム内のクロスリファレンス機能を提供する．ソースプログラムを XHTML でマークアップし，また各種識別子についての情報をまとめた情報テーブルという XML を出力することでハイパーリンクを利用したクロスリファレンスを Web ブラウザを用いて実現する．

表 1: ソースプログラムの class 属性が示す意味

属性値	意味
Func	関数名
Var	大域変数名
L_Var	局所変数名
Const	定数値
Arg	引数名
Tag	構造体タグ名
Member	構造体メンバ名
Macro	マクロ名

SPIE が出力する XHTML 化したソースプログラムでは識別子は<a>でマークアップされ，その class 属性には表 1 に挙げた値が指定されている．これによって識別子の種類が示される．本手法では，表 1 に挙げた値を class 属性に持つ要素を識別子として対応付けの対象とした．

3.2　ドキュメントの記述形式

ドキュメントは DocBook [3] と呼ばれる XML の形式で記述されたものを扱う．DocBook とはソフトウェアの関連文書向けの XML タグセットである．元々UNIX の文書交換のために作成され，現在も FDP[1]や LDP[2]で用いられている．またオープンソースソフトウェアのドキュメンテーションプロジェクトである OSWG[3]でも採

[1]FreeBSD Documentation Project
[2]Linux Documentation Project
[3]Open Source Writers Group

用されており，多くのソフトウェア関連文書の記述に用いられる形式であることから本研究でも対象として採用した．

DocBook にはプログラムに含まれる要素を示すためのタグが用意されているので，これを用いて識別子を得る．本手法において対応付けの対象となる要素を表すタグを表 2 に示す．

表 2: 識別子を示すタグ

タグ	意味
function	関数名
symbol	マクロ
literal	リテラル値
structname	構造体名
structfield	構造体のメンバ
type	型
varname	変数名

3.3 文脈情報の利用

ソースプログラム中の識別子の出現には「宣言」であるものと「参照」であるものの 2 種類がある．このことをドキュメントにも反映し，ドキュメント中の識別子の出現を「定義」と「参照」に区別する．DocBook の文脈を利用し，識別子の親要素に表 3 に挙げる要素を持つ場合に「定義」であると判断する．この表に挙げたタグは本質的には識別子の定義であることを示すタグではない．しかし，意味やその使用法を考慮して識別子の定義であることを示していると思われるものを選択した．

表 3: 識別子を定義と判断するタグ

タグ	意味
title	タイトル
grossterm	定義リストの項目名
funcprototype	関数のプロトタイプ

識別子の出現を「定義」と「参照」に区別することは，対応付けの応用において利用できる．例えば，ある関数名についてその意味を調べるためにドキュメントを参照する場合，「定義」の記述に意味が記述されている可能性が高い．また，ソースプログラムとドキュメントの整合性を検査する場合にもドキュメント中に「定義」の無い識別子のリストアップなどの詳細な検証を行うことができる．

3.4 対応付け

ソースプログラムとドキュメントのそれぞれに含まれる識別子の出現を種類と名前が一致するものを組み合わせて対応付ける．

対応付けを応用するためにはファイル中のそれぞれの識別子の出現が特定できなくてはならない．ソースプログラム中の識別子については SPIE が付与した要素の name 属性の値を用いる．この値は全体のプログラムでそれぞれがユニークである．一方，ドキュメント中の識別子は要素の id 属性を用いる．これはファイル中でユニークであることが定められているが，必須の値ではないので作成者が付与していない場合が多い．この場合，対応付けの前処理としてこの id 属性の補完が必要である．

対応付けの処理の流れを図 1 に示す．ソースプログラムを SPIE によりマークアップし，識別子を抽出する．ドキュメントは id 属性を補完して識別子を抽出する．それぞれから抽出した識別子の出現に対して対応付けを行い，対応表を出力する．対応表は XML ファイルで保存し，応用ツールが利用できるようにする．

3.5 評価

対応付けシステムを実装し，愛知県立大学情報科学部情報システム学科のプログラミング実習で作成するコンパイラとその 2002 年度版指導書を対象に対応付けを行い，検証した．

結果を表 4 に示す．誤った対応は引数，局所変数，定数についての対応であった．引数や局所変数の対応付けに誤りが多いのはスコープルールの問題によるものである．ソースプログラムではスコープ内で重複しなければ，異なる実体を示す同名の識別子を定義できる．しかしドキュメントから識別子を取り出す際にはこのスコープを考慮することができないので，同名

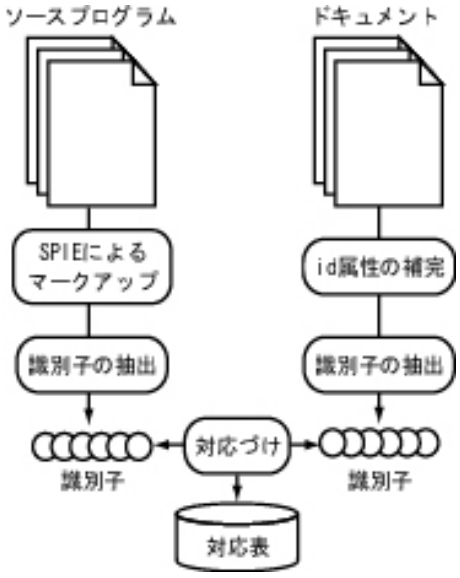

図 1: 対応付けの流れ

表 4: 対応付けの結果

ドキュメントのファイル数	8
行数	2643 行
ファイルサイズ	104KB
ソースプログラムのファイル数	5
行数	4667 行
ファイルサイズ	115KB
対応の数	1803
誤った対応	31

の識別子は全て同じ実体を示すものとして扱われてしまう．より正確な対応付けのためには解決しなくてはならない．

定数についても，識別子と同様の対応付けを試みたが，意味を考慮できないので正確な対応付けができなかった．識別子のようにある実体を指す名前ではなく，関係をその値で判断しなくてはならないので同じものが多数あり対応付けることが難しかった．重要な定数はマクロ定義で用いられることが多く，定数自体の対応付けは，利用者がオプションによって指定した場合のみ対応付けを行うようにすることとした．

4 対応付けの応用

3 章で行った対応付けを応用するツールとしてクロスリファレンスツールと整合性検査ツールを紹介する．

クロスリファレンスツールではソースプログラムとドキュメントの相互参照機能を提供し，ソフトウェアの理解を支援する．

整合性検査ツールはドキュメントの妥当性を検証して校正作業をしたり，ソースプログラムを仕様書との比較によって作成したプログラムが仕様に則っているのかを検証したりすることに利用できる．

4.1 クロスリファレンスツール

クロスリファレンスツールは対応付けに基づいたソースプログラムとドキュメントの間での相互参照機能を提供するツールである．SPIE がソースプログラム内のクロスリファレンスを提供するために出力するファイル群に，ドキュメントの相互参照をリンクとして埋め込む．SPIE はソースプログラムを XHTML でマークアップすると共に，各種識別子についての情報をまとめた情報テーブルを XML ファイルとして出力する．この情報テーブルに，それぞれのドキュメントでの識別子が出現へのリンクを列挙することでソースプログラムからドキュメントへの参照を可能にする．またドキュメント上の識別子の出現には，対応付けられた識別子の情報テーブルへのリンクを付加する．このように，情報テーブルを介してソースプログラムとドキュメントの相互参照を実現する．

識別子のドキュメント上の出現箇所を列挙する際，3.3 節で示した識別子の「定義」と「参照」の区別によって表示方法を変化させる．特に優先的に参照したいであろう「定義」の記述へのリンクは強調表示をする．これによって利用者はリストの中から定義へのリンクを選択することができる．

Web ブラウザを用いた実際の画面を図 2 に掲載する．図中左下のウィンドウには情報テーブルが表示され，右上のウィンドウにドキュメントが表示されている．右上のウィンドウはソースプログラムとドキュメントの兼用で，情報

テーブルで指定されたものが表示される．

図 2: クロスリファレンス画面

4.2 評価

3.5 節と同様にプログラミング実習の教材を利用して機能を検証した．実際にクロスリファレンス機能を提供するファイル群を生成し，その講義を受講したことのある学生 5 人に使用した感想を聞き，以下のような回答を得た．

- 実際の指導書は内容が前後している部分があり，必要な箇所を探すのに手間取ることが多かったが，このツールを使えばすぐに見つけられた．
- 前回の内容を参照しなくてはならないときに，忘れてしまっていてもすぐに指導書から見つけられた．

4.3 整合性検査ツール

このツールは，対応付けに基づいてこの整合性についての検査を行うツールである．ソースプログラムとドキュメントの間での識別子の存在に注目した整合性検査を行う．ソースプログラムのみに含まれる識別子とドキュメントのみに含まれる識別子をリストアップする．前者は主にドキュメント上での書き漏れが原因と考えられる．ソースプログラムの修正が先行し，ドキュメントの修正が行われていない場合に起こりやすい．後者は例えば仕様書で決められた識別子名が実際には用いられていない場合などに検出される．仕様にそっていないため，ソースプログラムの不具合の発見にも役立つと考えられる．

また，3.3 節で示した「定義」と「参照」の区別により，ドキュメント上に「定義」のない識別子をリストアップすることも可能である．

4.4 評価

3.5 節同様，プログラミング実習の教材を用いてツールの検証を行った．図 5 のような結果を得た．

表 5: 整合性検査の結果

ソースプログラム中の識別子数	534
ドキュメント中の識別子数	99
対応の数	1804
ソースプログラム中のみの識別子	451
ドキュメント中のみの識別子	41

ソースプログラムでは 84%の識別子が，ドキュメントでは 41%の識別子が不整合として検出された．不整合の割合が大きいのはこの場合のドキュメントが指導書であり，その性質上プログラムの詳細を全て記述する必要はなく，省かれているものが多くあったためにソースプログラムのみに出現する識別子が多くあったと考察した．また，指導書には作成したプログラム以外のプログラムの解説も掲載されており，このことによってドキュメントのみに出現した識別子があった．この結果から，ドキュメントとして指定するファイルは仕様書などのソースプログラムの詳細を記したものが適当であるといえる．

リストアップされた識別子を観察した結果，プログラミング実習の指導書として記載されているべき識別子が記載されていなかった例を関数名で 3 件，マクロで 1 件発見できた．従って，これらはドキュメントの修正すべき点と判断でき，このツールの有効性を確認できた．

現在は単純に対応付けられているかどうかの検査のみであるので，誤記などを検出することは不可能である．近くに記述されている識別子や文脈などの情報を利用して，より高度な整合性の検査を実現すればより有用なツールになる．

5 おわりに

5.1 まとめ

本稿ではソースプログラムとドキュメントのそれぞれに含まれる識別子の出現を対応付け，その結果を応用する手法を提案した．

対応付けの際，ドキュメント中の識別子の抽出時に，その出現文脈から「定義」と「参照」を区別する方法を用いた．このことによって識別子名の一致のみによる対応付けよりも有用なものになった．

対応付けを行うプロセスとその結果を応用するプロセスを分けたことでそれぞれにまだ改良を施すことができる．対応付けをある特定の目的のために行わないので，本稿で紹介した応用ツールのほかにもさまざまな用途に利用できる．

5.2 今後の課題

現段階では応用ツールを利用するためには対応付けを行う必要がある．従って，それぞれの編集に対してインタラクティブな機能を提供することはできない．ソースプログラムを作成中にドキュメントの関係箇所を表示する機能や，ドキュメントを誤りを直しながら校正する作業のためには，こうした変更に対して毎回の対応付けを必要としない方法が必要である．

本稿で実装した対応付けシステムではドキュメント上の識別子に関してはスコープルールを考慮できない．ドキュメント上の識別子についてもソースプログラム同様のスコープルールが適用でき，異なる実体を示す同名の識別子が存在しうる．しかし，現在の対応付け手法では同名の識別子は全て同じものとして扱われてしまう問題がある．そこで，局所変数については近くにある関数名などを調べることによってどの関数内で用いられる変数なのかを判定できると考えられる．文脈や近くにある要素などによって意味を考慮した対応付けを行い，より多くの意味を持った対応付けを行う必要がある．

謝辞

ご指導，ご議論頂いた愛知県立大学情報科学部 稲垣康善教授，名古屋大学大学院情報科学研究科 阿草清滋教授，および山本研究室の皆様に感謝します．本研究は文部科学省科学研究費補助金基盤研究 (A)『「ソフトウェア=プログラム+ドキュメント」の視点に基づく多言語対応大規模コーパス』(課題番号 16200001) を受けて行われた．

参考文献

[1] Sapid, http://www.sapid.org

[2] 福安 直樹, 山本 晋一郎, 阿草 清滋, “細粒度ソフトウェア・リポジトリに基づいた CASE ツール・プラットフォーム Sapid”, 情報処理学会論文誌, Vol.39, No.6, pp.1990–1998, 1998

[3] OASIS, http://www.oasis-open.org

[4] 大橋 洋貴，山本 晋一郎，阿草 清滋，“ハイパーテキストに基づいたソースプログラム・レビュー支援ツール”，電子情報通信学会技術研究報告，Vol.98, No.28，pp.15–22，1998

[5] 大場勝， 権藤克彦， “アスペクト指向を用いたドキュメント整理法の提案”， 日本ソフトウェア科学会 第 7 回プログラミングおよび応用のシステムに関するワークショップ， 2004, http://spa.jssst.or.jp/2004/pub/papers/04027.pdf

[6] 大場勝， “OS 入門用の教材を事例としたアスペクト指向によるソースコードとドキュメントの関連づけ”， 北陸先端科学技術大学院大学情報科学研究科 修士論文， 2004, http://www.jaist.ac.jp/library/thesis/is-master-2004/paper/m-ohba/paper.ps

[7] 山田信幸, 鈴木幹雄, 坂守,“XML 形式の仕様書作成によるソフトウェア机上チェックの効率化－机上デバッグ支援ツール Prio（プライオ）－”，ソフトウェアテストシンポジウム 2004，ソフトウェアテストシンポジウム 2004 予稿集，pp.92–99, http://www.swtest.jp/, 2004

[8] D.E.Knuth, 有澤誠 訳, “文芸的プログラミング”, アスキー出版局, 1994